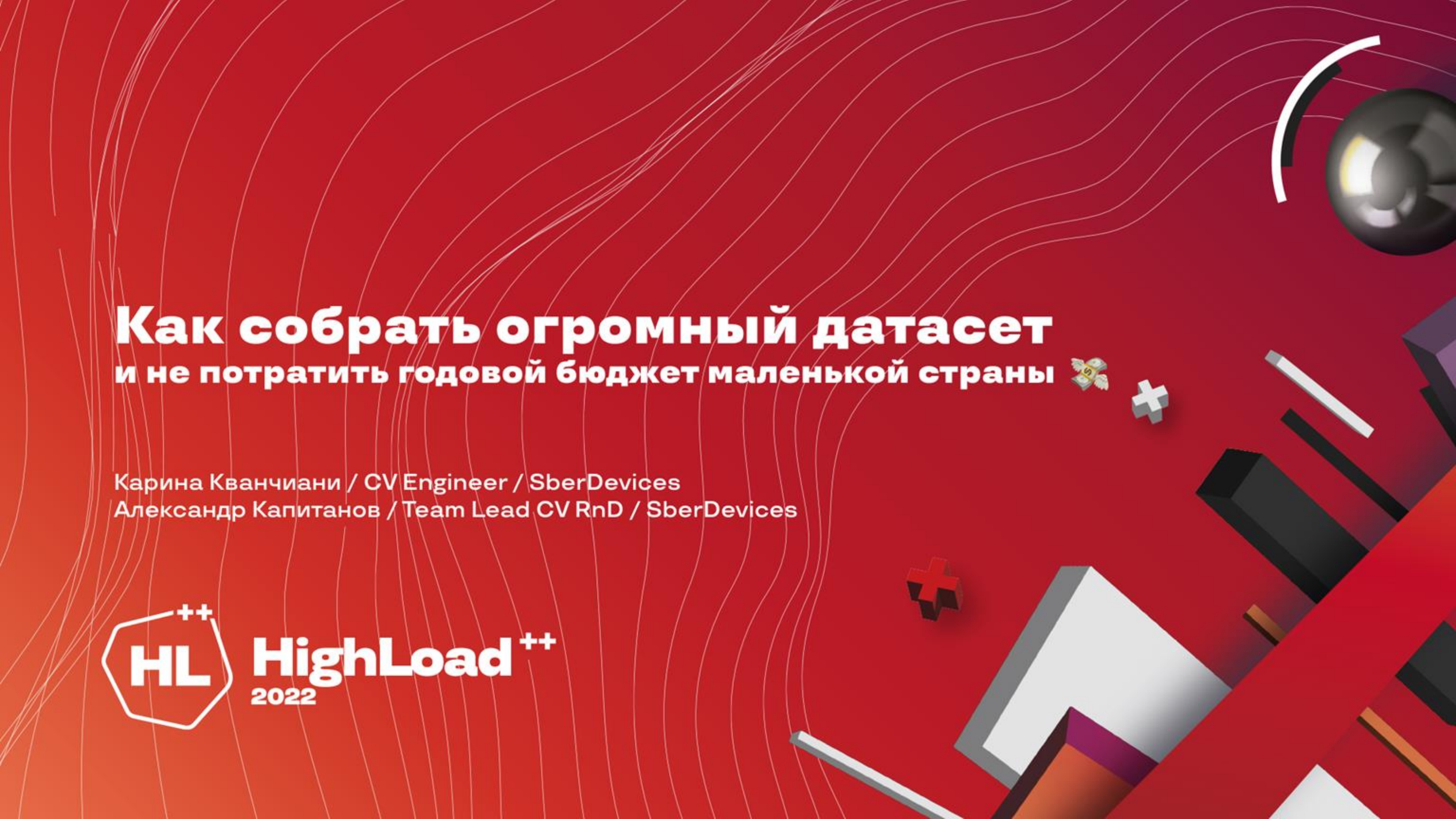


Как собрать огромный датасет
и не потратить годовой бюджет маленькой страны
Карина Кванчиани / CV Engineer / SberDevices
Александр Капитанов / Team Lead CV RnD / SberDevices
HL
HighLoad++
2022



Как собрать огромный датасет
и не потратить годовой бюджет маленькой страны 💸

Карина Кванчиани / CV Engineer / SberDevices
Александр Капитанов / Team Lead CV RnD / SberDevices

Девайсы и Jazz

SBER DEVICES

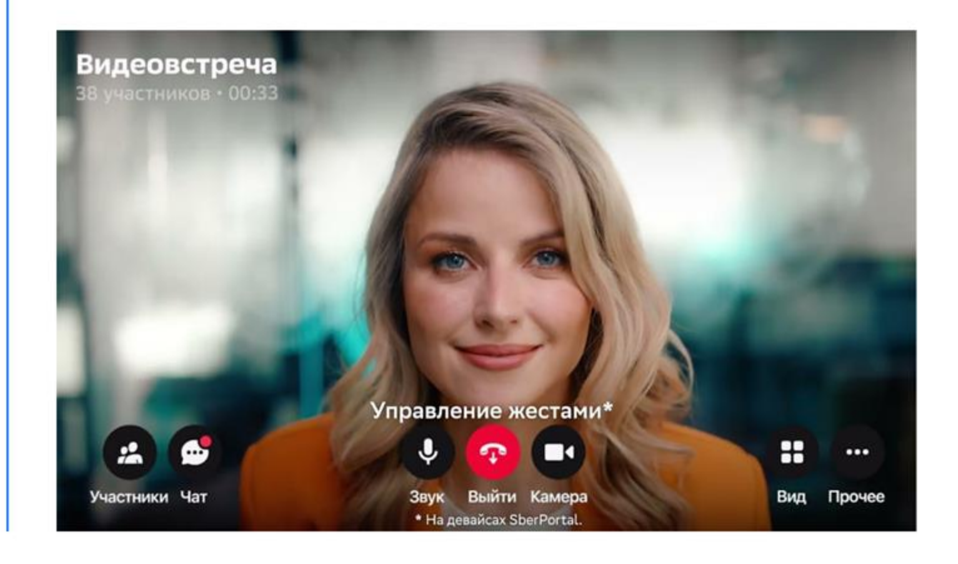


Салют

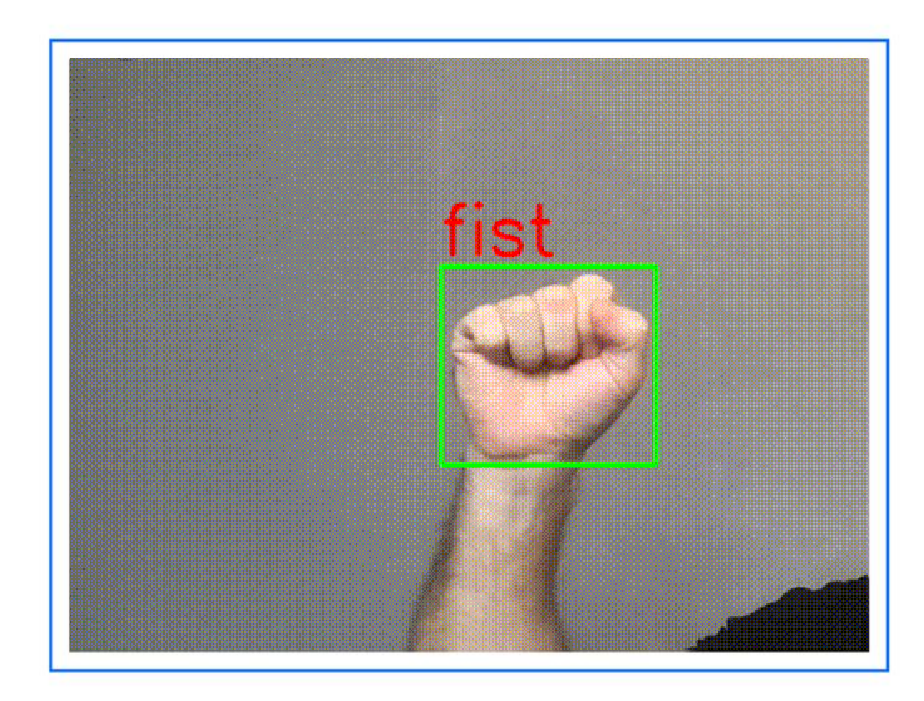




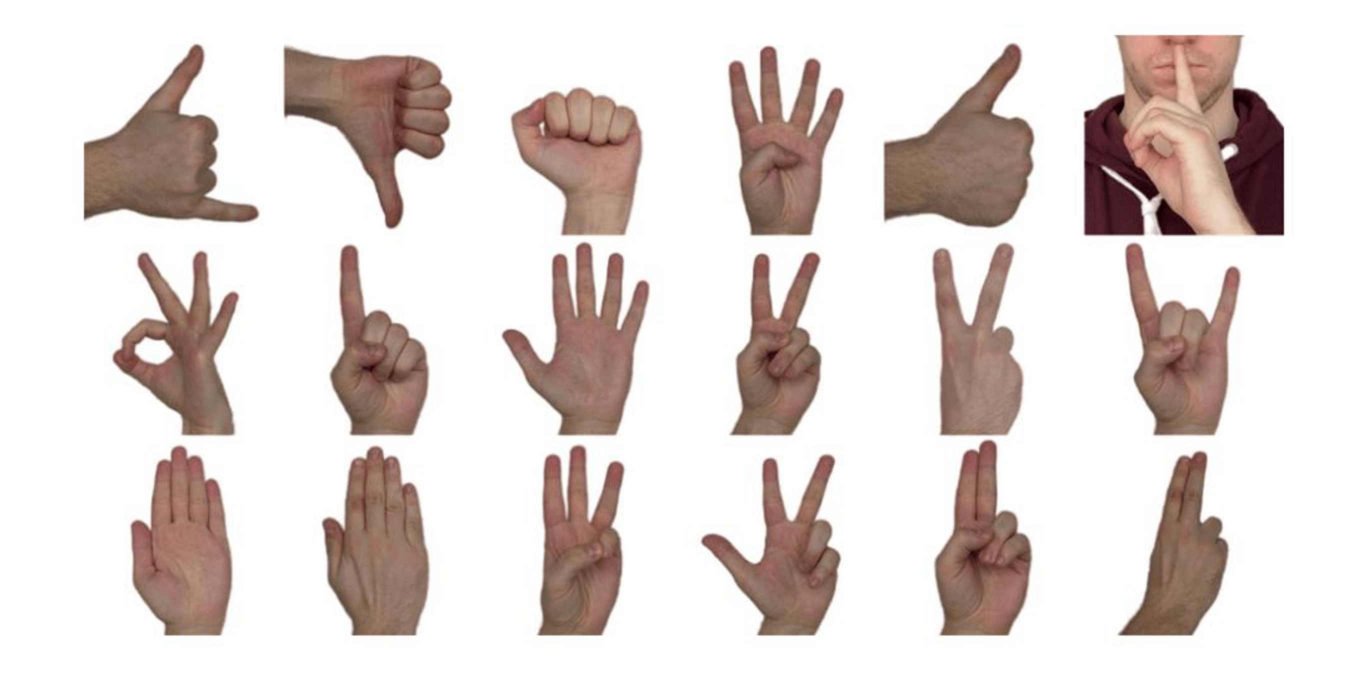
Jazz by Sber



Распознавание жестов и данные



Распознавание жестов и данные



HaGRID

HAnd Gesture Recognition Image Dataset



552.992
изображений
в FullHD

18 классов
жестов

дополнительный
класс “не жест”

34,730
различных субъектов

>= 34,730 🤟
различных сцен

от 18 до 65
возраст субъектов

ж / м = 27 / 20
пол субъектов

искусственный и
естественный свет +
экстремальные условия

от 0.5 до 4 метров
расстояние
субъектов до камеры

План доклада

SBER DEVICES

- Краудсорсинг-платформы
- Пайплайн сбора и разметки данных
- Минимизация затрат и повышение качества разметки
- Публикация в open-source
- Автоматизация пайплайна

Краудсорсинг-платформы



Yandex.Toloka

готовый продукт
лучше интерфейс
больше разметчиков



ABC Elementary

есть модератор
“идеальные” разметчики
кастомные проекты

Краудсорсинг-платформы

проекты, пулы и как это все работает

заказчик

↘

проект
разметить фото

→

пул 1
обучение

пул 2
экзамен

пул 3
основные задания 1

пул 4
основные задания 2

...

→

задание 1
n заданий

задание 2
n заданий

...

↘

исполнитель

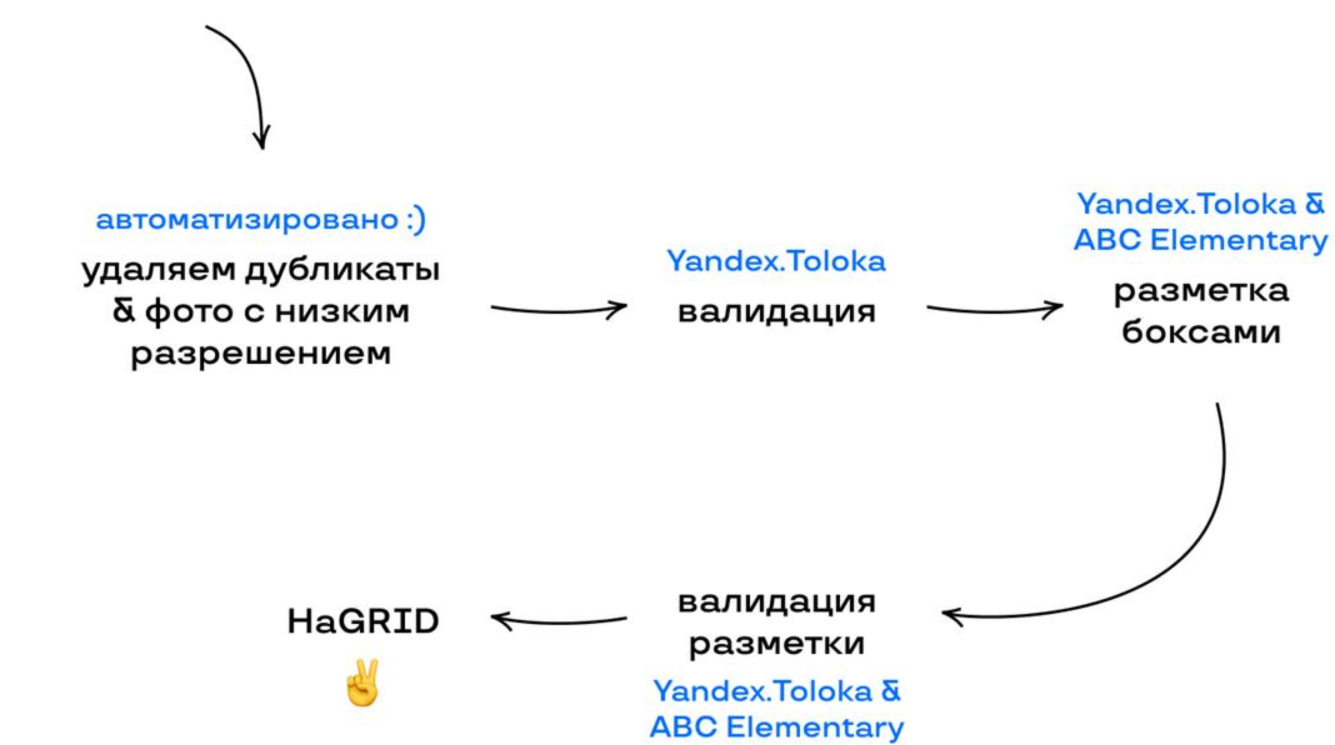




Пайплайн сбора и разметки данных

SBER DEVICES

Yandex.Toloka
сбор изображений



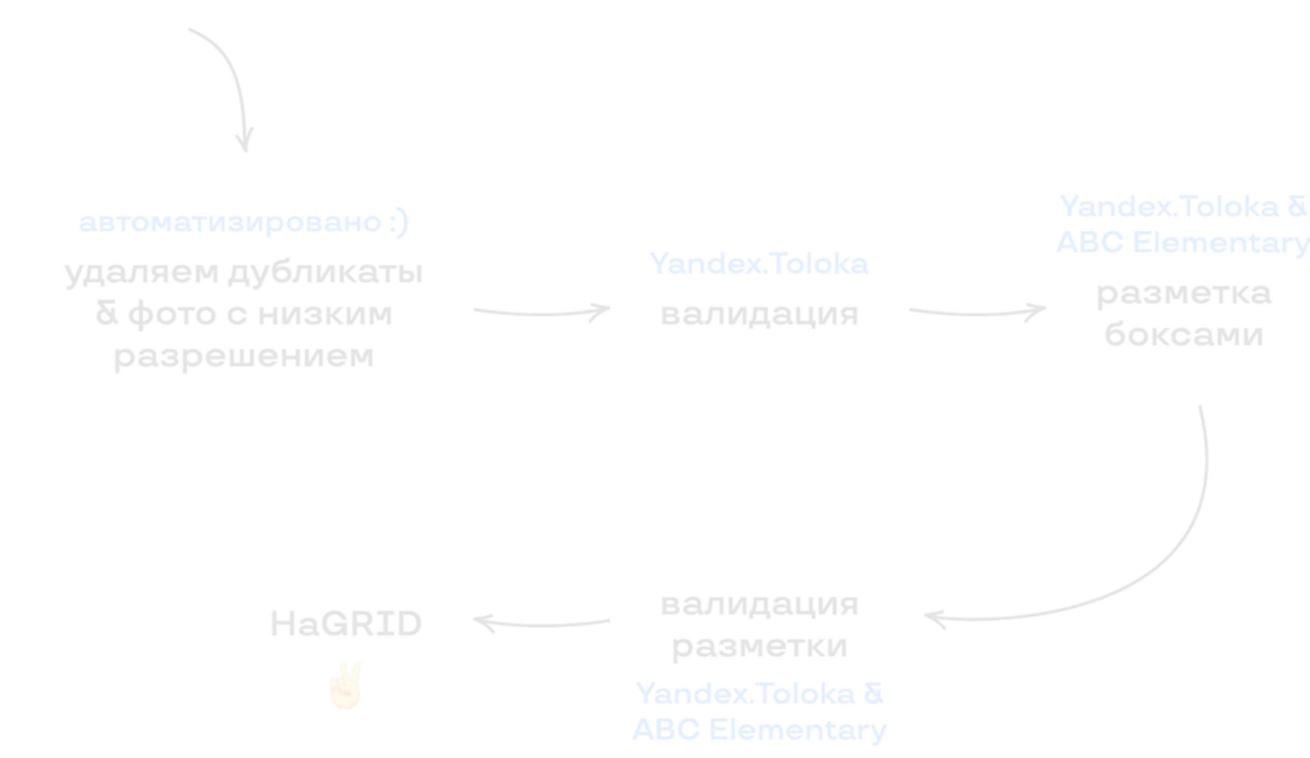
мы делаем здесь amazing customer service

Пайплайн сбора и разметки данных

SBER DEVICES

Yandex.Toloka

сбор изображений



Пайплайн сбора и разметки данных
сбор изображений

SBER DEVICES

Ваше фото с жестом:

Жест виктори (указательный и средний палец образуют знак V)

Шаблон к заданию



Ваше фото

Сделать фото

Задача:

показать указанный в задании жест и сделать фото себя

Условия: ✌️

рука с жестом должна быть в кадре полностью

фото должно быть сделано с расстояния от 1 до 4 метров

Пайплайн сбора и разметки данных
сбор изображений

Соотношение скорости и качества

Обратите внимание, чем меньше исполнителей выбрано, тем ниже скорость выполнения пула

Подробнее

ТОП % | Онлайн

Пул будет доступен указанному проценту исполнителей с лучшим рейтингом



Мы отобрали **80%** лучших исполнителей

Задание доступно **8647** активным исполнителям

Пайплайн сбора и разметки данных

сбор изображений

Цена

Цена за страницу заданий, $ *

Рекомендованная цена для простых заданий — 0,02 $

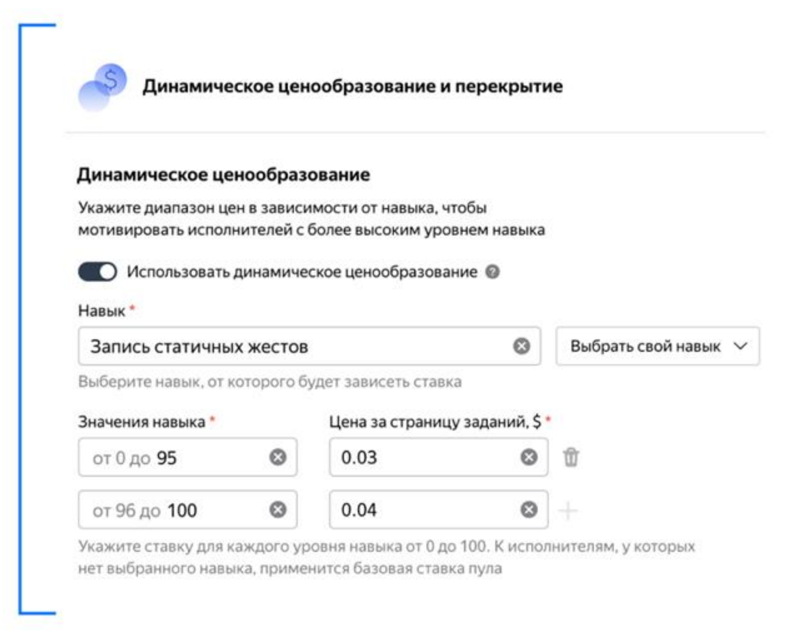
Рекомендованная цена для сложных заданий — 0,05 $

0.01

Динамическое ценообразование за страницу заданий

для навыка Запись статичных жестов

0,03 $ – 0,04 $



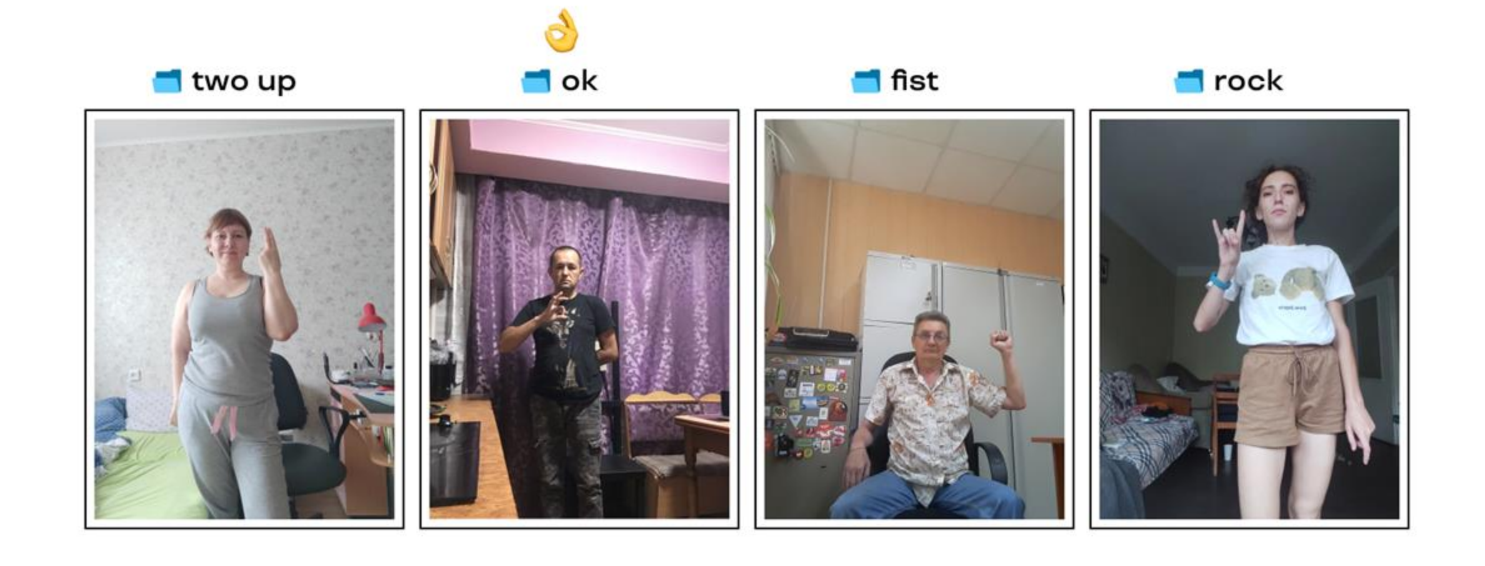
Пайплайн сбора и разметки данных

сбор изображений

Название	Приоритет	Прогресс	Статус	Запущен	Будет завершён
CALL [RNDCV] [Alexander Kapitanov]	0	8005* / 10000	Идёт разметка...	18 октября 2022 г., 9:22	~ 2 часа
LIKE [RNDCV] [Alexander Kapitanov]	0	6143* / 8000	Идёт разметка...	18 октября 2022 г., 9:21	~ час
DISLIKE [RNDCV] [Alexander Kapitanov]	0	6134* / 8000	Идёт разметка...	18 октября 2022 г., 9:21	~ 2 часа
FIST [RNDCV] [Alexander Kapitanov]	0	6785* / 8000	Идёт разметка...	18 октября 2022 г., 9:21	~ час
STOP [RNDCV] [Alexander Kapitanov]	0	6645* / 9000	Идёт разметка...	18 октября 2022 г., 22:56	~ 11 часов
STOP_INVERTED [RNDCV] [Alexander Kapitanov]	0	1765* / 2000	Идёт разметка...	13 октября 2022 г., 13:20	~ 20 часов
THREE [RNDCV] [Alexander Kapitanov]	0	6273* / 8000	Идёт разметка...	18 октября 2022 г., 9:21	~ 5 минут
TWO_UP [RNDCV] [Alexander Kapitanov]	0	712* / 1000	Идёт разметка...	18 октября 2022 г., 21:58	~ 8 часов
THREE2 [RNDCV] [Alexander Kapitanov]	0	681* / 1000	Идёт разметка...	18 октября 2022 г., 22:22	~ 9 часов
FOUR [RNDCV] [Alexander Kapitanov]	0	11000 / 11000	Размечен	9 октября 2022 г., 19:28	—
PALM [RNDCV] [Alexander Kapitanov]	0	11000 / 11000	Размечен	28 сентября 2022 г., 15:48	—
MUTE [RNDCV] [Alexander Kapitanov]	0	11000 / 11000	Размечен	7 октября 2022 г., 17:05	—

Пайплайн сбора и разметки данных

сбор изображений



Пайплайн сбора и разметки данных

Yandex.Toloka
сбор изображений

автоматизировано :)
удаляем дубликаты
& фото с низким
разрешением

Yandex.Toloka
валидация

Yandex.Toloka &
ABC Elementary
разметка
боксами

валидация
разметки
Yandex.Toloka &
ABC Elementary

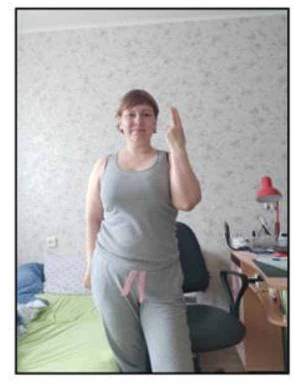
HaGRID
✌️

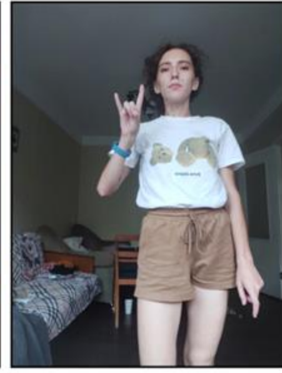
SBER DEVICES

Пайплайн сбора и разметки данных

удаляем дубликаты & фото с низким разрешением

SBER DEVICES





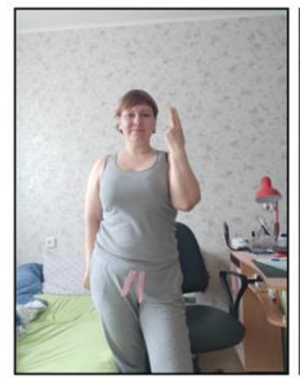
Пайплайн сбора и разметки данных

удаляем дубликаты & фото с низким разрешением

SBER DEVICES

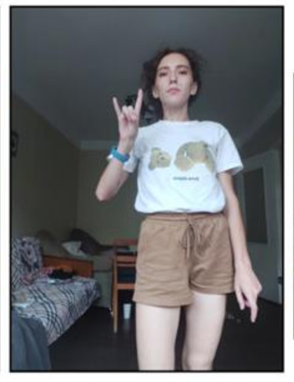


шаблон













дубликат

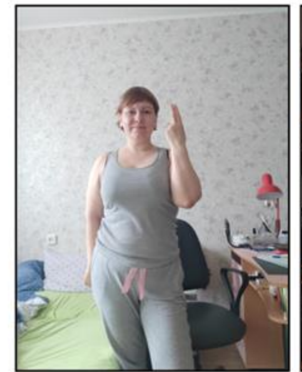
низкое разрешение

Пайплайн сбора и разметки данных

удаляем дубликаты & фото с низким разрешением

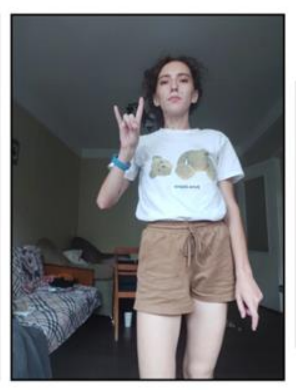
SBER DEVICES



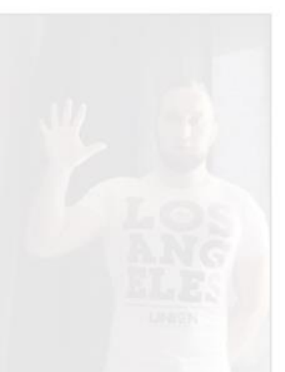












дубликат

низкое разрешение

Пайплайн сбора и разметки данных

уровни недобросовестных исполнителей

SBER DEVICES

0 lvl

дубликат





Пайплайн сбора и разметки данных

уровни недобросовестных исполнителей

0 lvl

дубликат





5 lvl

шаблон



Пайплайн сбора и разметки данных

уровни недобросовестных исполнителей

0 lvl

дубликат





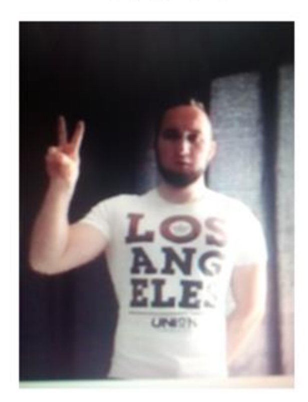
5 lvl

шаблон



20 lvl

шаблон



с обработкой

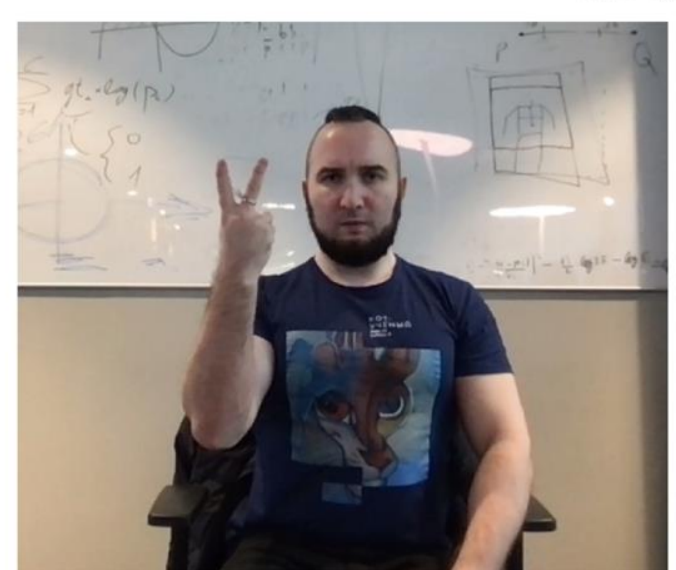
Пайплайн сбора и разметки данных

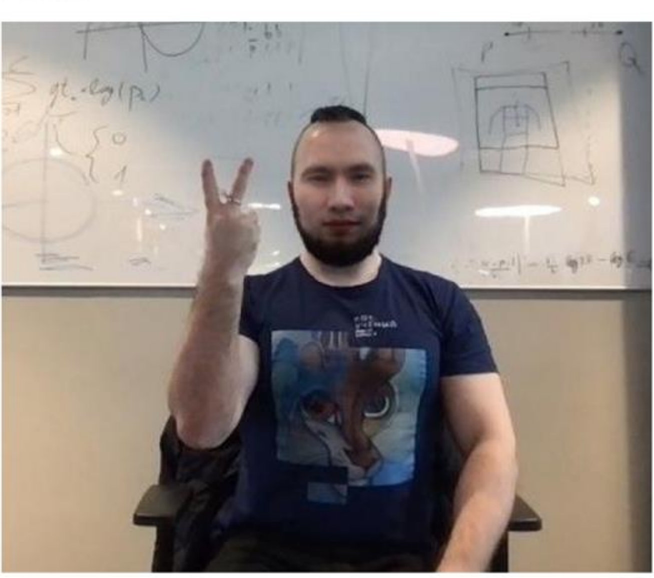
уровни недобросовестных исполнителей

SBER DEVICES

80 lvl

дипфейки





просто fun to remember



Пайплайн сбора и разметки данных

SBER DEVICES

Yandex.Toloka
сбор изображений

автоматизировано :)
удаляем дубликаты & фото с низким разрешением

Yandex.Toloka
валидация

Yandex.Toloka & ABC Elementary
разметка боксами

валидация разметки
Yandex.Toloka & ABC Elementary

HaGRID
✌️

Пайплайн сбора и разметки данных

валидация

Жест кулак (рука сжата в кулак) с расстояния 3-4 метра от камеры

Выберите подходящий вариант ответа:

1 Правильно

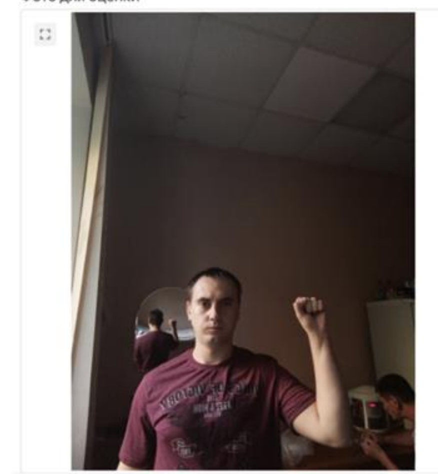
2 Неправильно

3 Фото повернуто

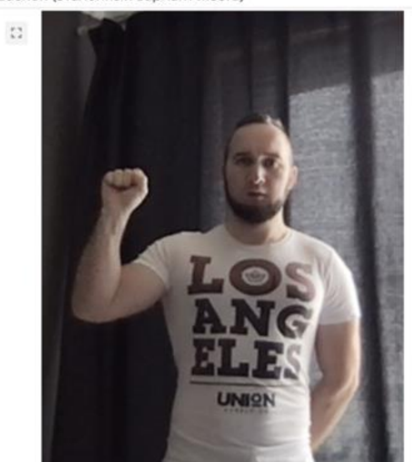
4 Стороннее фото

5 Не загрузилось

Фото для оценки







Задача:

определить, верно ли выполнен жест на фото в соответствии с шаблоном

Критерии выбора:

жест показан верно

фото / жест не соответствует критериям

изображение повернуто

стороннее фото из интернета

не загрузилось

Пайплайн сбора и разметки данных

валидация

пул

обучение

100% лучших пользователей
(отобраны Я.Толокой)

10 заданий с подсказкой

навык обучения
% верных ответов

0 $

Пайплайн сбора и разметки данных

валидация

пул обучение	🤞 пул экзамен
100% лучших пользователей (отобраны Я.Толокой)	пользователи с навыком обучения 70% и выше
10 заданий с подсказкой	10 заданий без подсказки
навык обучения % верных ответов	навык экзамена % верных ответов
0 $	0 $

Пайплайн сбора и разметки данных

валидация

SBER DEVICES

пул обучение	🤞 пул экзамен	пул основной
100% лучших пользователей (отобраны Я.Толокой)	пользователи с навыком обучения 70% и выше	пользователи с навыком экзамена 70% и выше
10 заданий с подсказкой	10 заданий без подсказки	10% контрольных заданий
навык обучения % верных ответов	навык экзамена % верных ответов	3 разметки на каждое изображение
0 $	0 $	0.01 $ / 0.06 $ / 0.09 $

Пайплайн сбора и разметки данных

SBER DEVICES

Yandex.Toloka
сбор изображений

автоматизировано :)
удаляем дубликаты & фото с низким разрешением

Yandex.Toloka
валидация

Yandex.Toloka & ABC Elementary
разметка боксами

валидация разметки
Yandex.Toloka & ABC Elementary

HaGRID
✌️

Пайплайн сбора и разметки данных

разметка боксами

SBER DEVICES

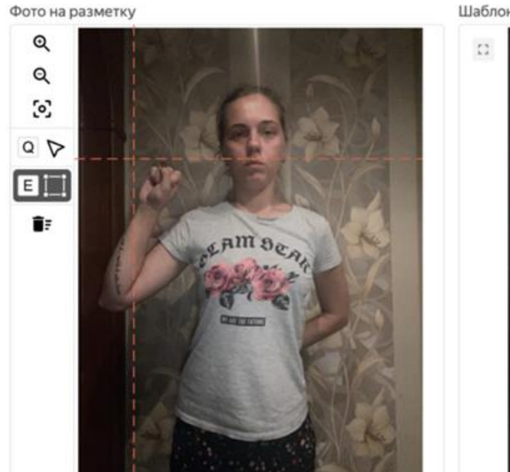
Условия:

- **ЕСЛИ КИСТЬ ОБРЕЗАНА - ОТМЕЧАТЬ НЕ НУЖНО!**
- Не забывайте указывать метки **Жест** и **Не жест**
- Сторонние фото, отсутствие кистей - помечаем чекбоксом

Жест кулак (рука сжата в кулак) с расстояния 3-4 метра от камеры

☐ Нет кистей / стороннее фото / повернуто

Фото на разметку



Шаблон (эталонный вариант жеста)



1 Жест | 2 Не жест

Задача:

выделить кисть руки с жестом красным прямоугольником, а без жеста – зеленым

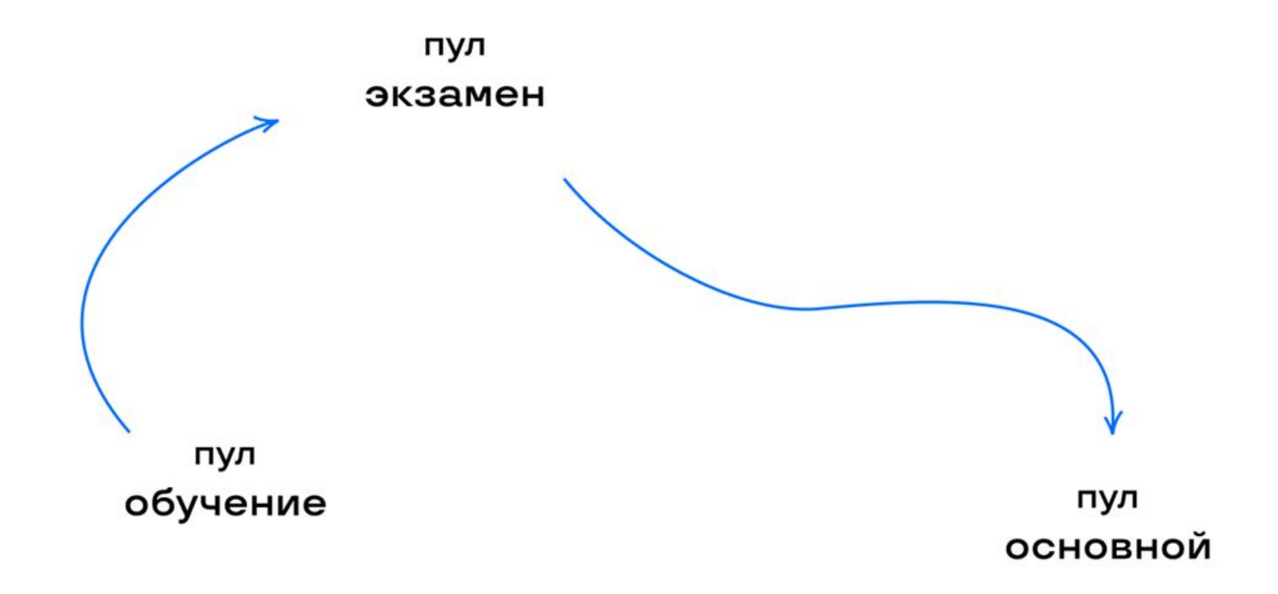
Условия: ✊

не выделять кисть, если она обрезана

каждому из боксов необходимо сопоставить метку – “жест” или “не жест”

Пайплайн сбора и разметки данных

разметка боксами



Пайплайн сбора и разметки данных

валидация разметки

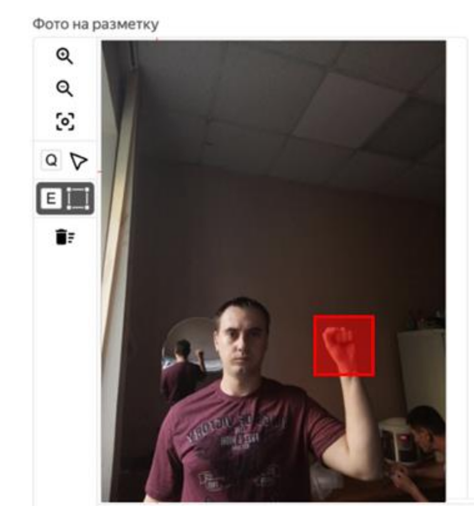
Условия:

- **ЕСЛИ КИСТЬ ОБРЕЗАНА - ОТМЕЧАТЬ НЕ НУЖНО!**
- Не забывайте указывать метки **Жест** и **Не жест**
- Сторонние фото, отсутствие кистей - помечаем чекбоксом

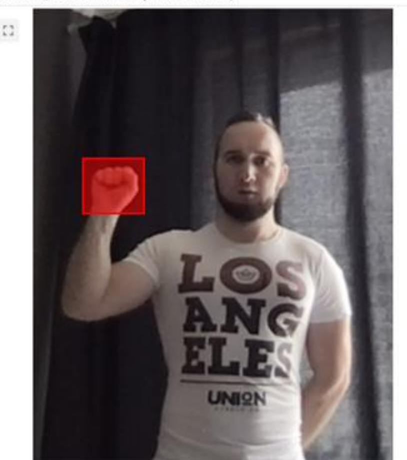
Жест кулак (рука сжата в кулак) с расстояния 3-4 метра от камеры

Нет кистей / стороннее фото / повернуто

Фото на разметку



Шаблон (эталонный вариант жеста)



1 Жест | 2 Не жест

Задача:

определить, верно ли выделены кисти рук на фото

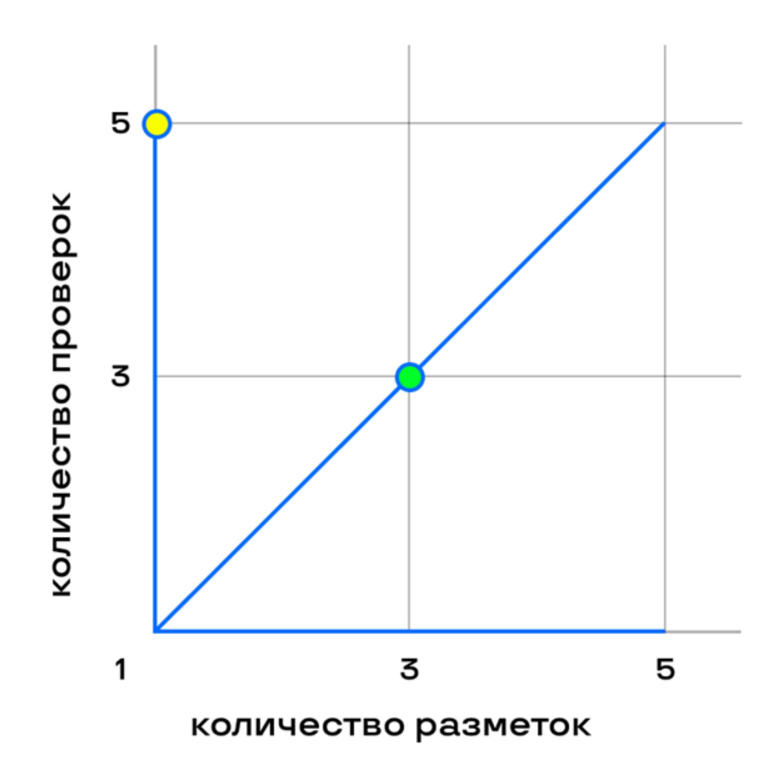
Критерии выбора:

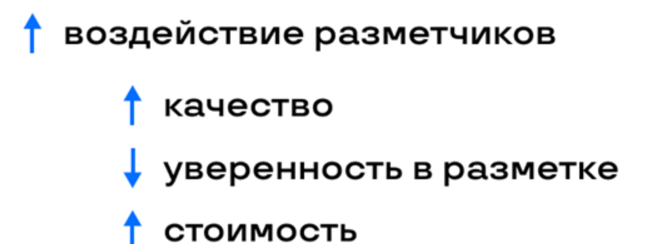
все кисти четко очерчены боксами

каждый из боксов имеет правильную метку – “жест” или “не жест”

Пайплайн сбора и разметки данных

разметка боксами и валидация разметки

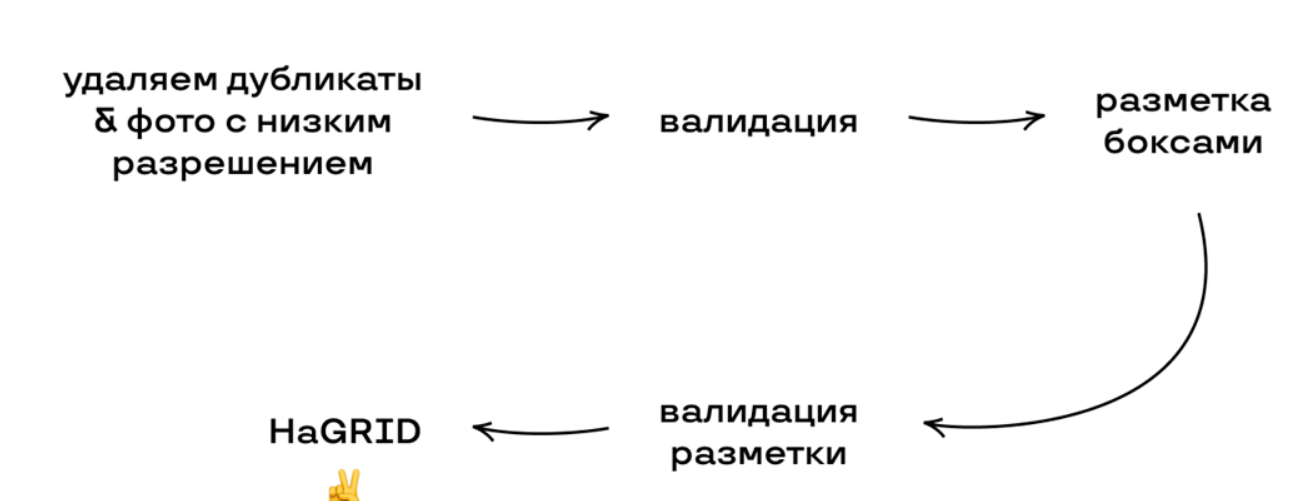




Пайплайн сбора и разметки данных

сбор изображений





SBER DEVICES

Стоимость пайплайна as is

каждое задание оценивалось в один цент (s = 0.01$)

	n	по этапам	1 изображение
сбор изображений	878.910	n × s ≈ 9.000$	0.01$
валидация	830.679	n × s × 3 ≈ 25.000$	s × 3 ≈ 0.03$
разметка	560.230	n × s × 5 ≈ 28.000$	s × 5 ≈ 0.05$
валидация разметки	560.230	n × s × 5 × 5 ≈ 140.000$	s × 5 × 5 ≈ 0.25$
классификация	552.992	n × s × 3 ≈ 17.000$	s × 3 ≈ 0.03$
итого	552.992	?	0.37$

Стоимость пайплайна as is



220.000$

BMW 8 серии Cabrio

или

GEFORCE RTX 4090

90 x



или

SberPortal

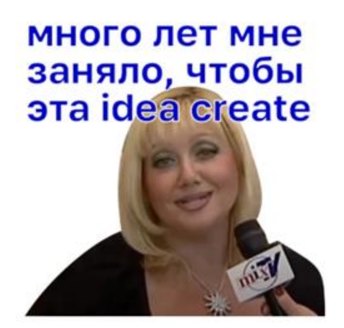
440 x



SBER DEVICES



дорого
&
некачественно



валидации разметки

может, ее автоматизировать?

Автоматизация валидации разметки

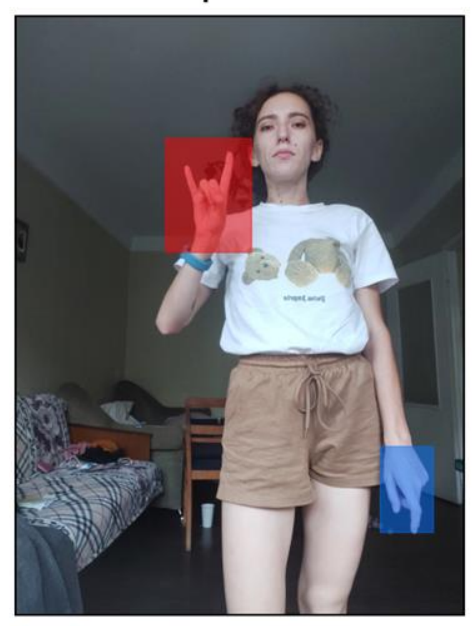
а какая бывает разметка?



идеальная



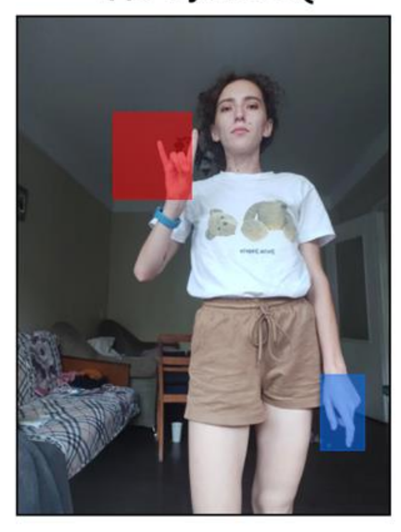
хорошая



Автоматизация валидации разметки

а какая бывает разметка?

бокс не там,
где нужно :(



Автоматизация валидации разметки

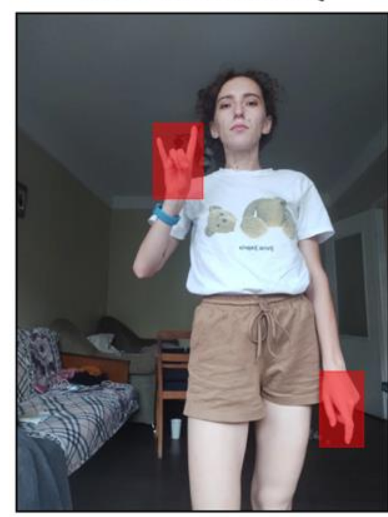
а какая бывает разметка?

SBER DEVICES

бокс не там,
где нужно :(



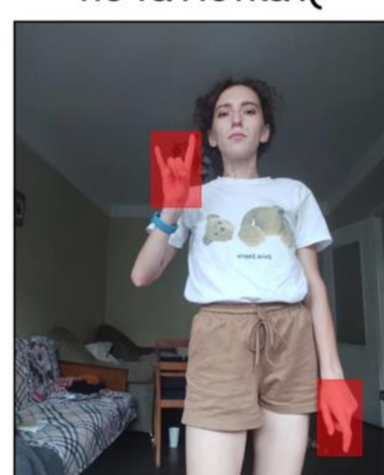
👎
не та метка :(



Автоматизация валидации разметки

а какая бывает разметка?

бокс не там, где нужно :(



не та метка :(



забыли бокс :(



Автоматизация валидации разметки

а какая бывает разметка?

SBER DEVICES





бокс не там,
где нужно :(

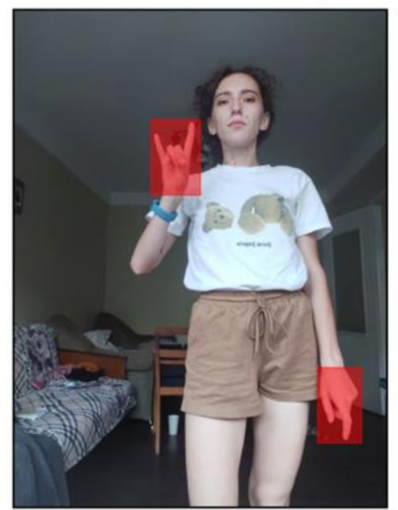
не та метка :(

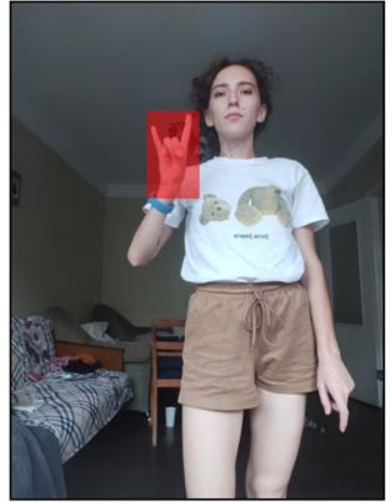
забыли бокс :(

нуу..





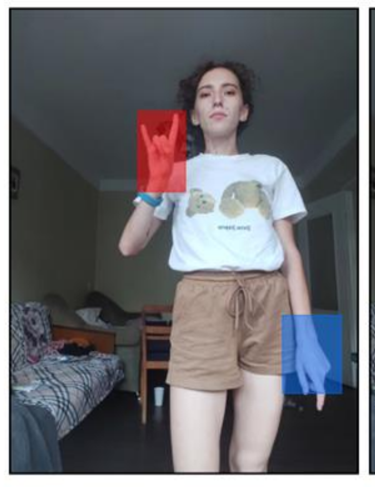


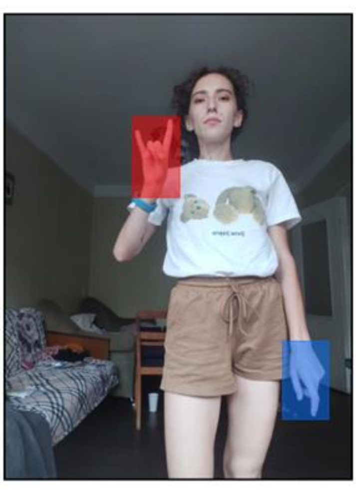


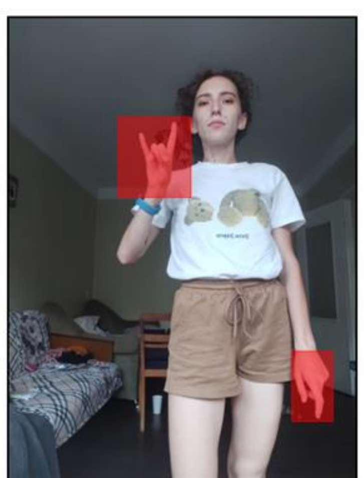


Агрегация разметки

3 разметки







Агрегация разметки



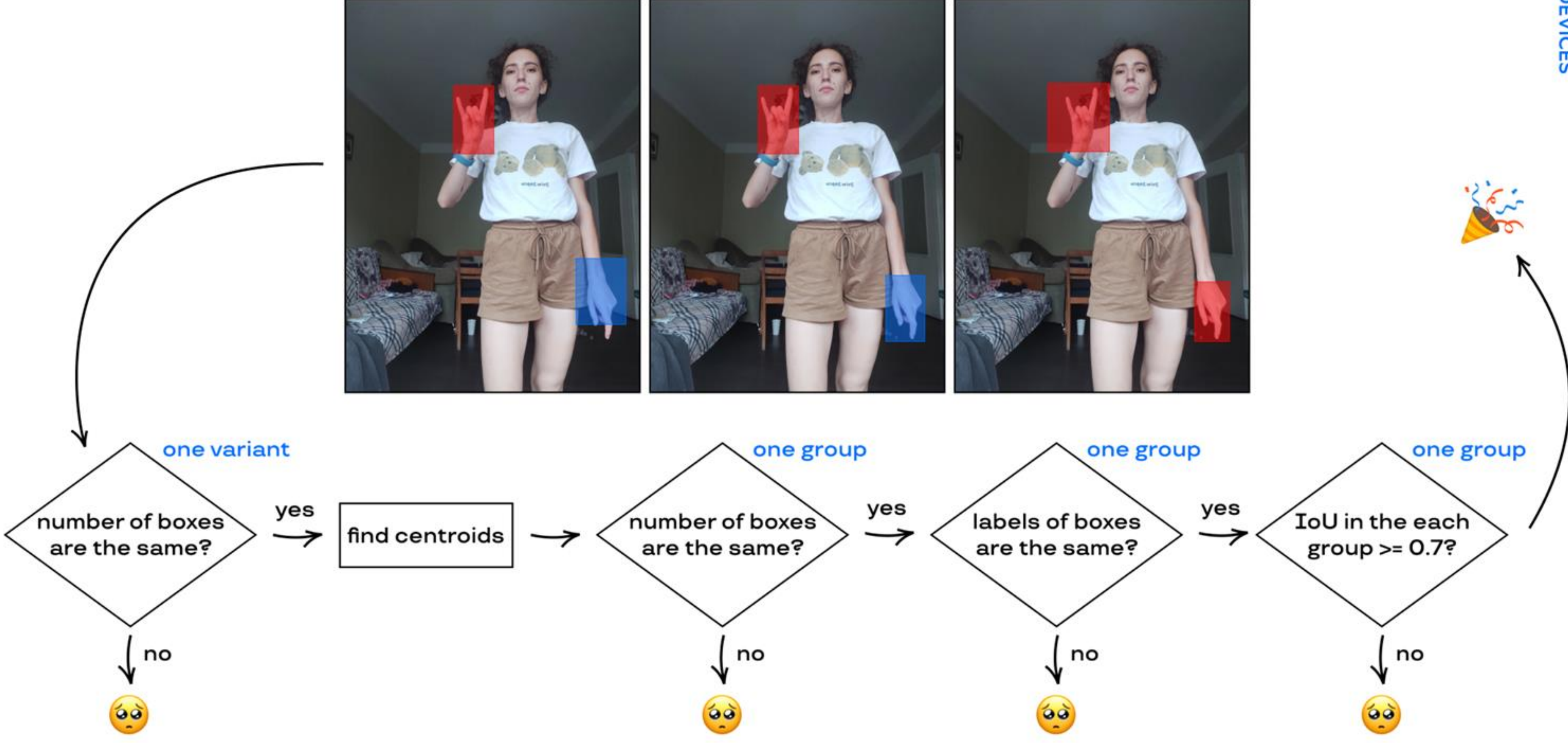
3 разметки



hard

Агрегация разметки

hard-подход



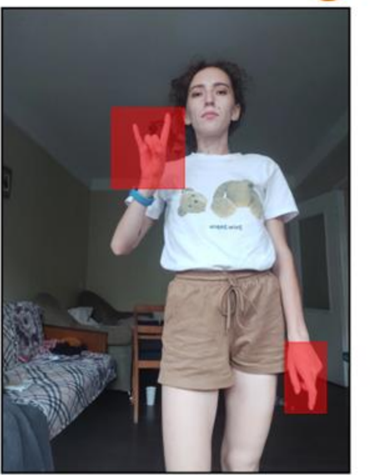
Агрегация разметки

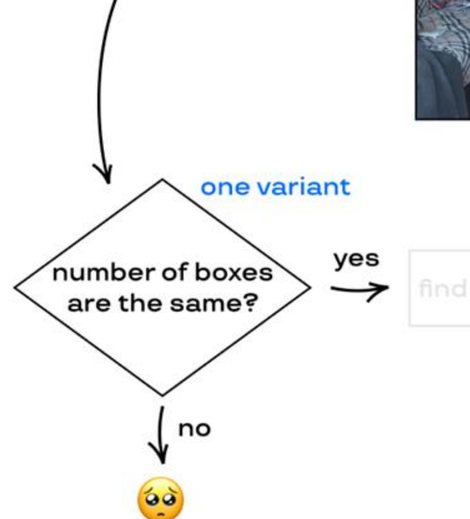
hard-подход

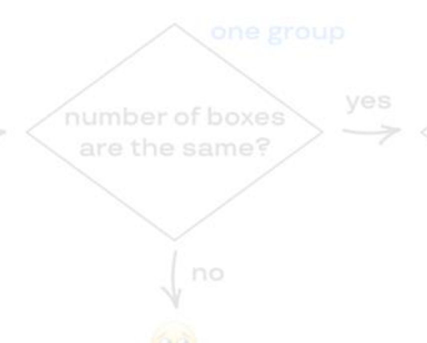
во всех разметках по 2 бокса 

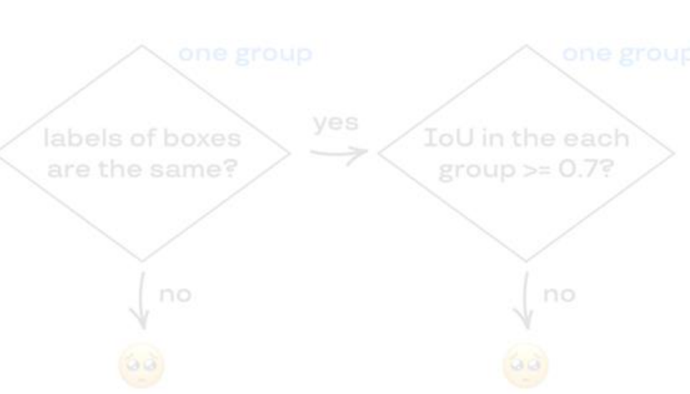












Агрегация разметки

hard-подход





one variant

number of boxes are the same?

yes → find centroids →

one group

number of boxes are the same?

yes →

one group

labels of boxes are the same?

yes →

one group

IoU in the each group >= 0.7?

no

no

no

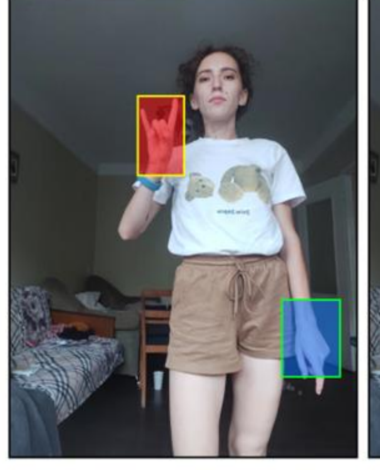
no

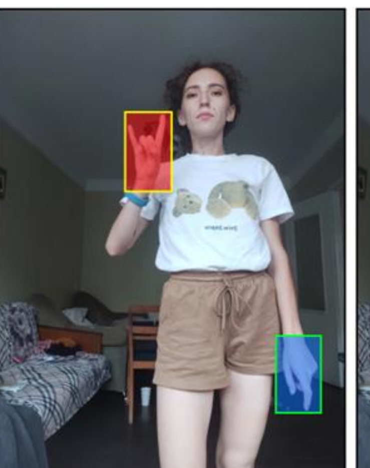
SBER DEVICES

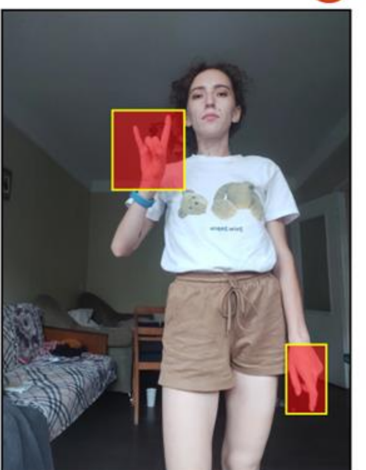
Агрегация разметки

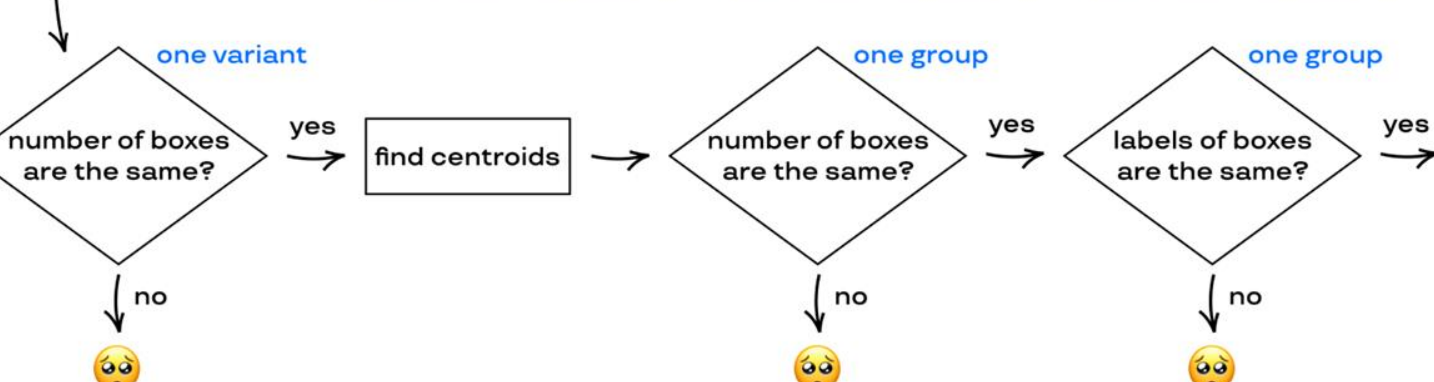
hard-подход

в одной разметке неверно указана метка 









Агрегация разметки

hard-подход

IoU в каждой из групп больше 70% 

one variant

number of boxes are the same?

yes → find centroids →

one group

number of boxes are the same?

yes →

one group

labels of boxes are the same?

yes →

one group

IoU in the each group >= 0.7?

no

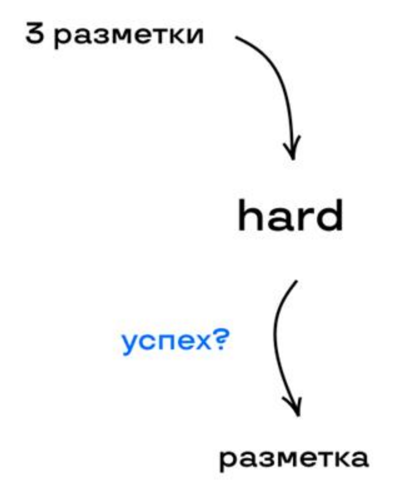
no

no

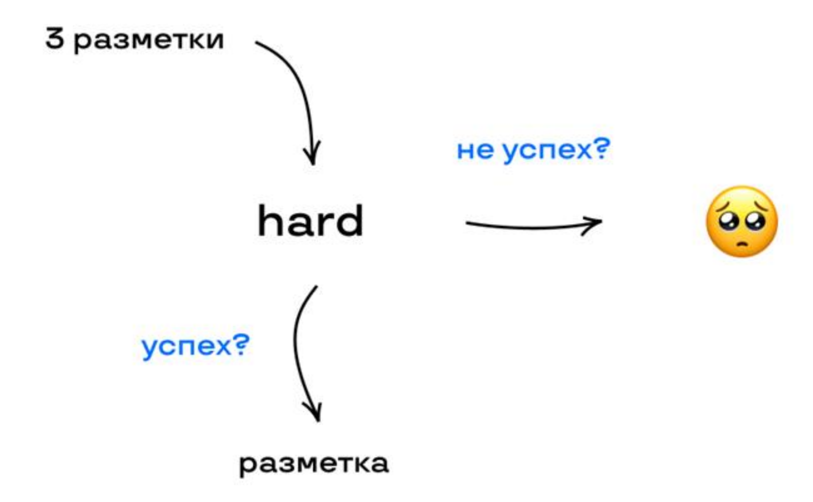
no

SBER DEVICES

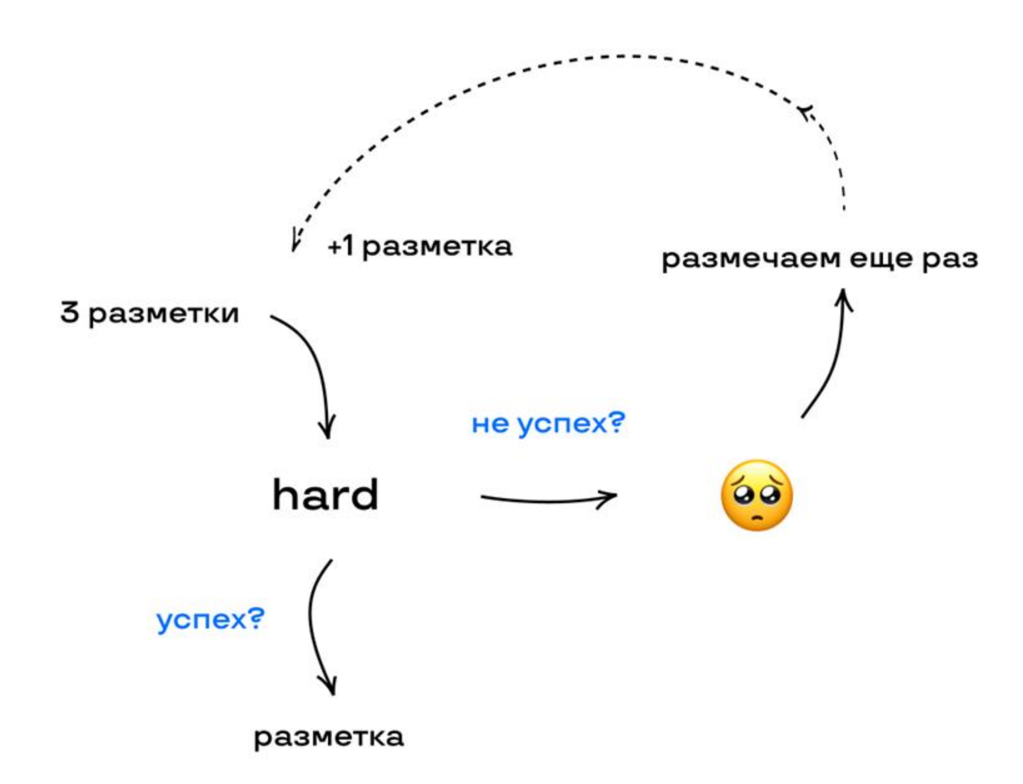
Агрегация разметки



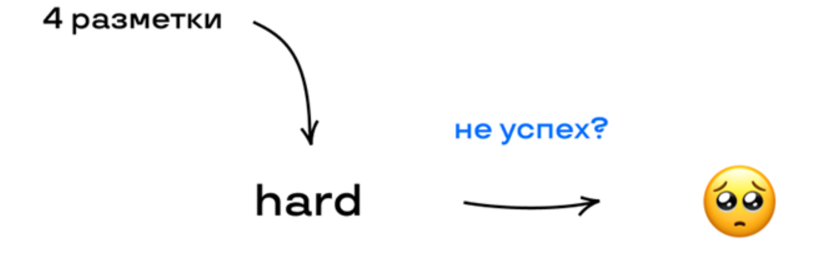
Агрегация разметки



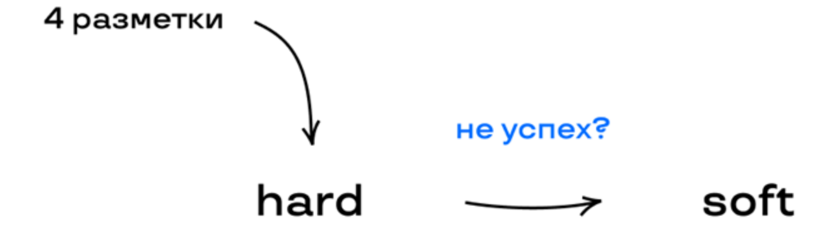
Агрегация разметки



Агрегация разметки

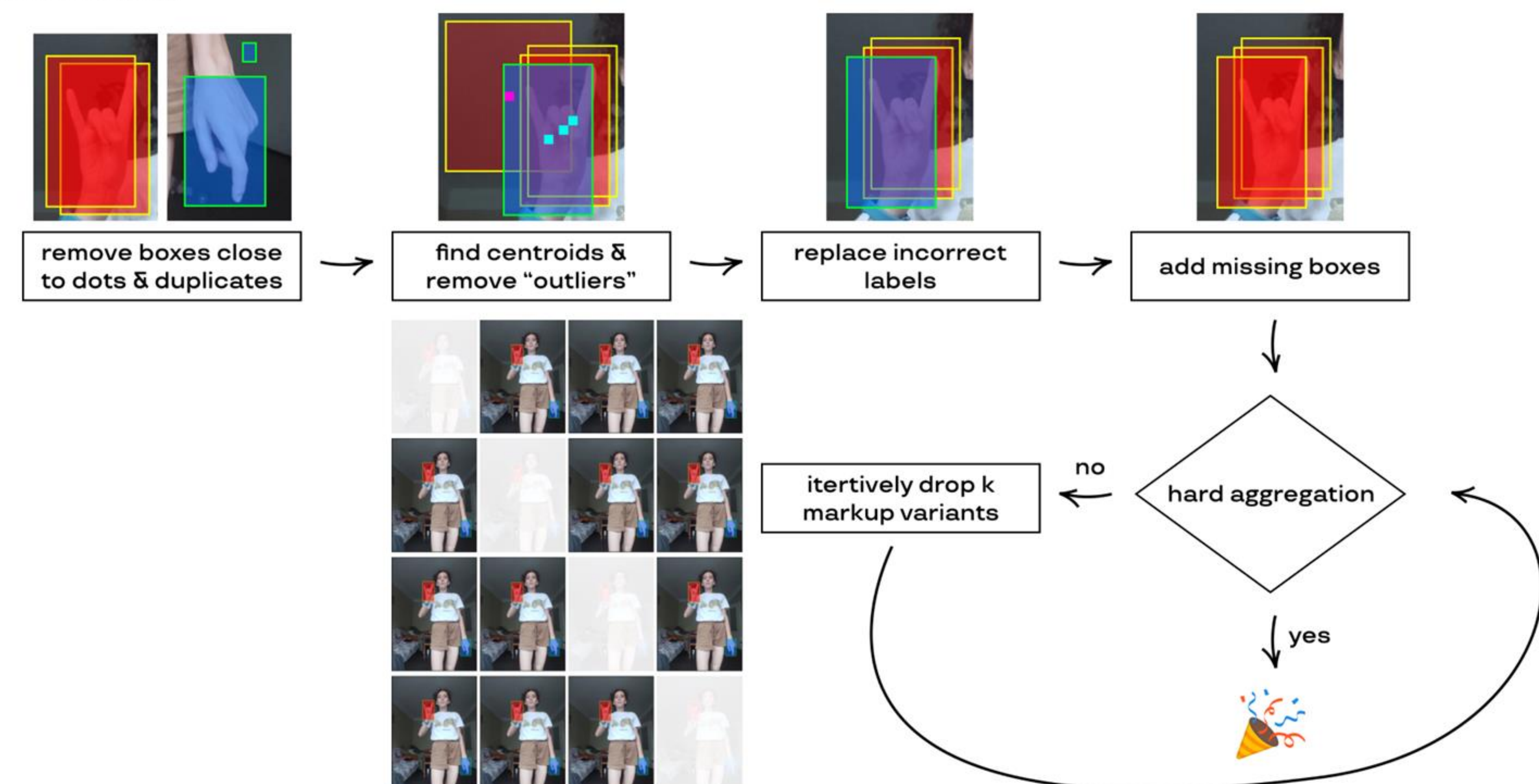


Агрегация разметки



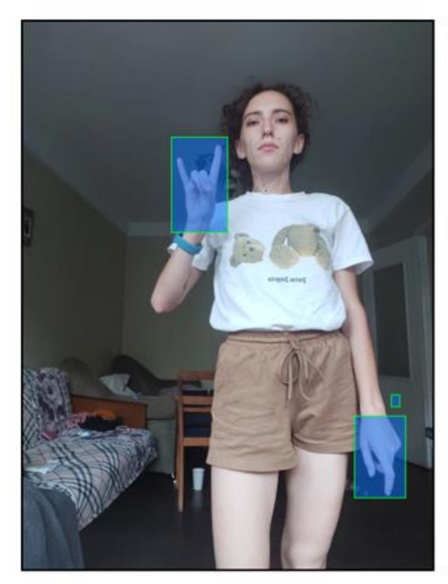
Агрегация разметки

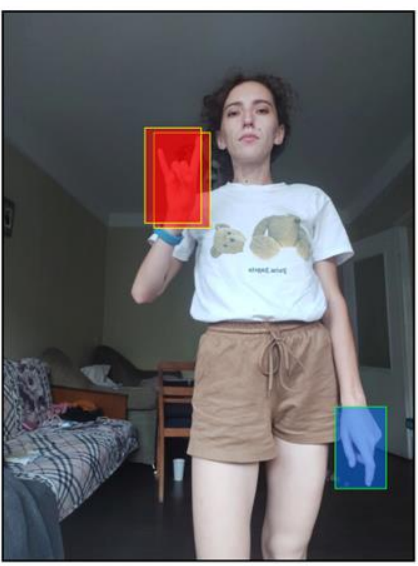
soft-подход

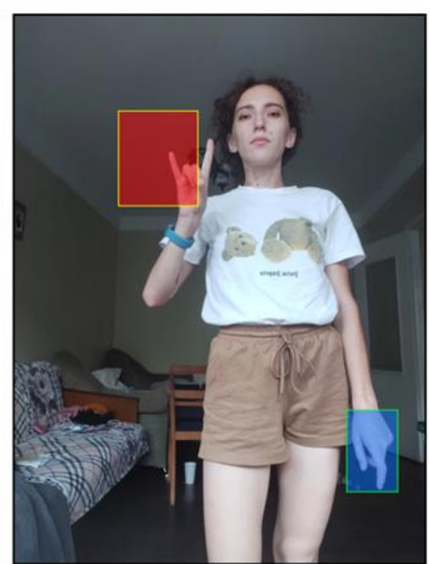


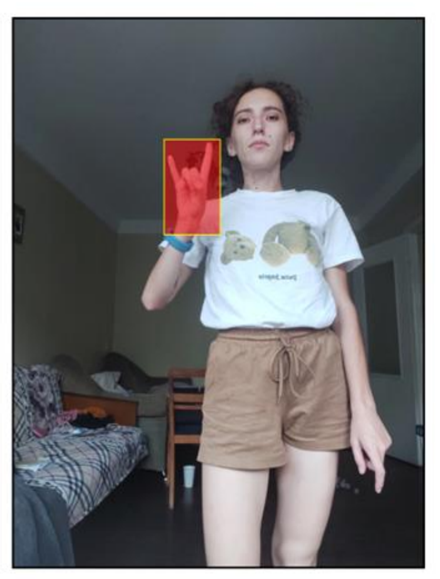
Агрегация разметки

soft-подход



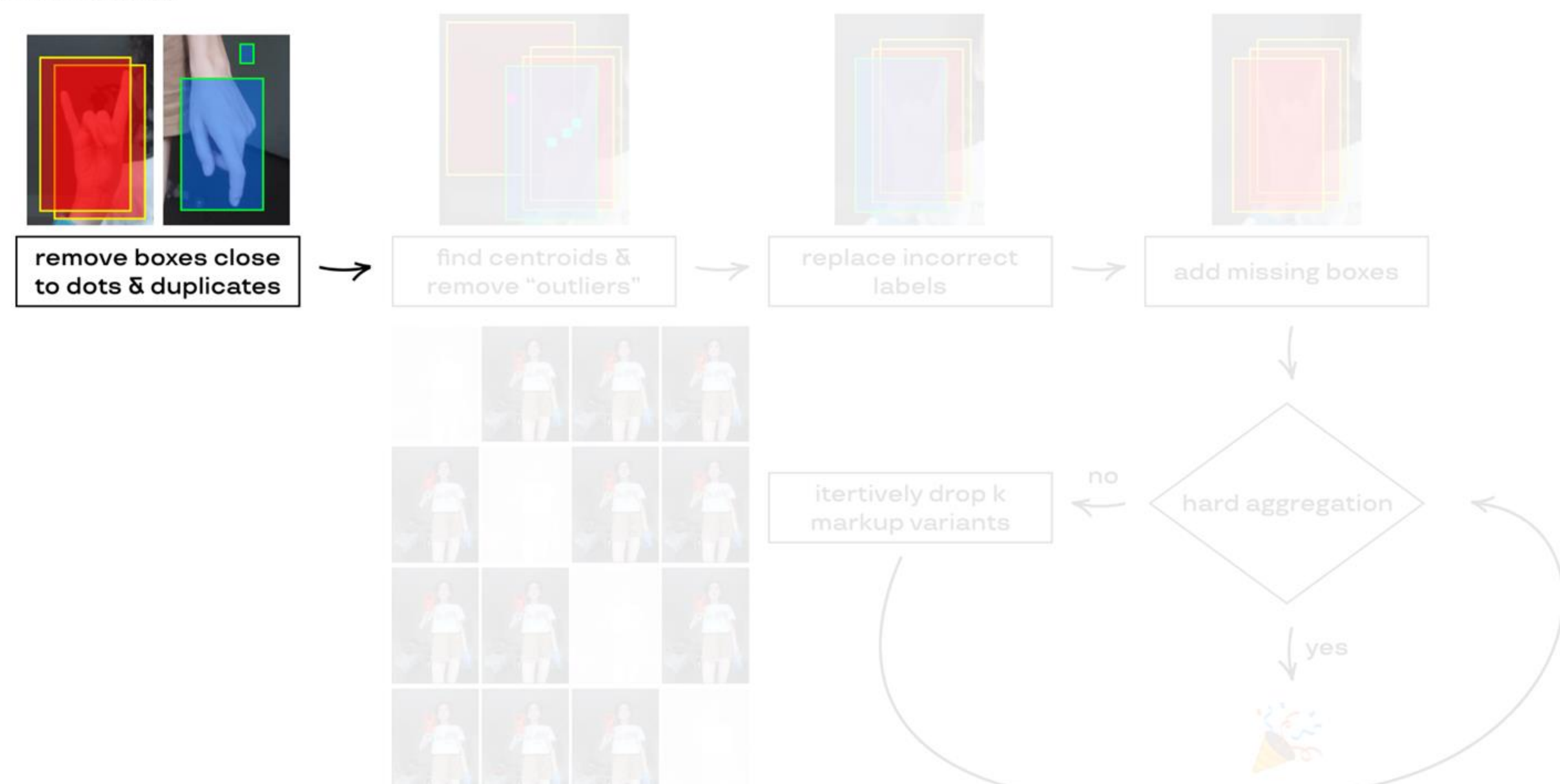






Агрегация разметки

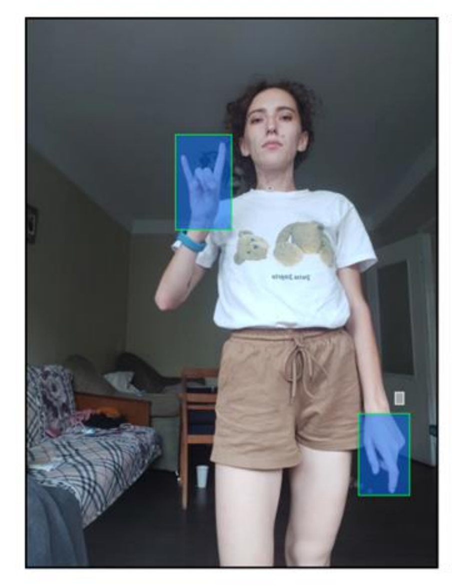
soft-подход

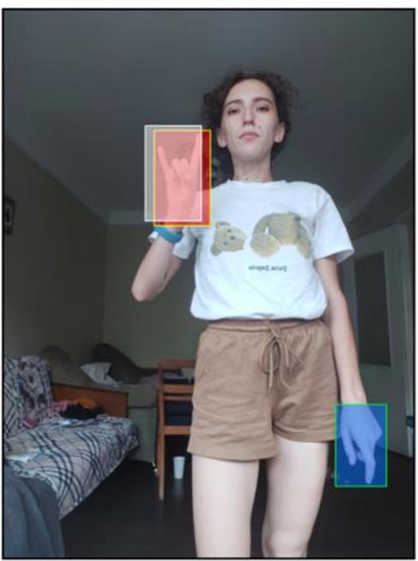


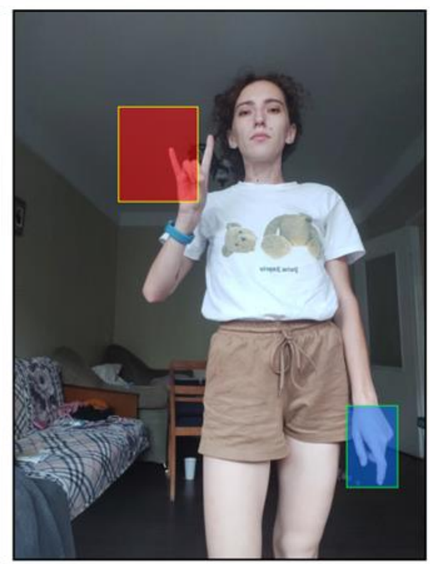
SBER DEVICES

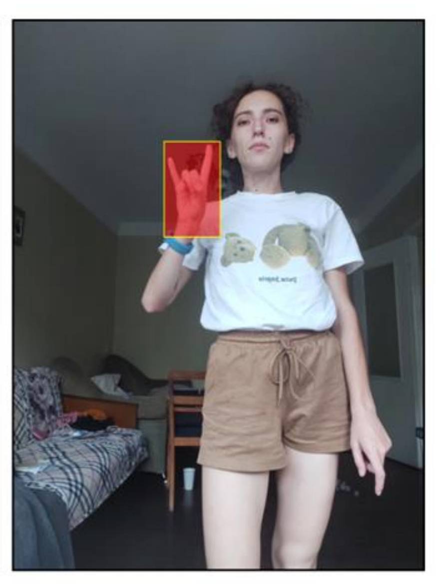
Агрегация разметки

soft-подход



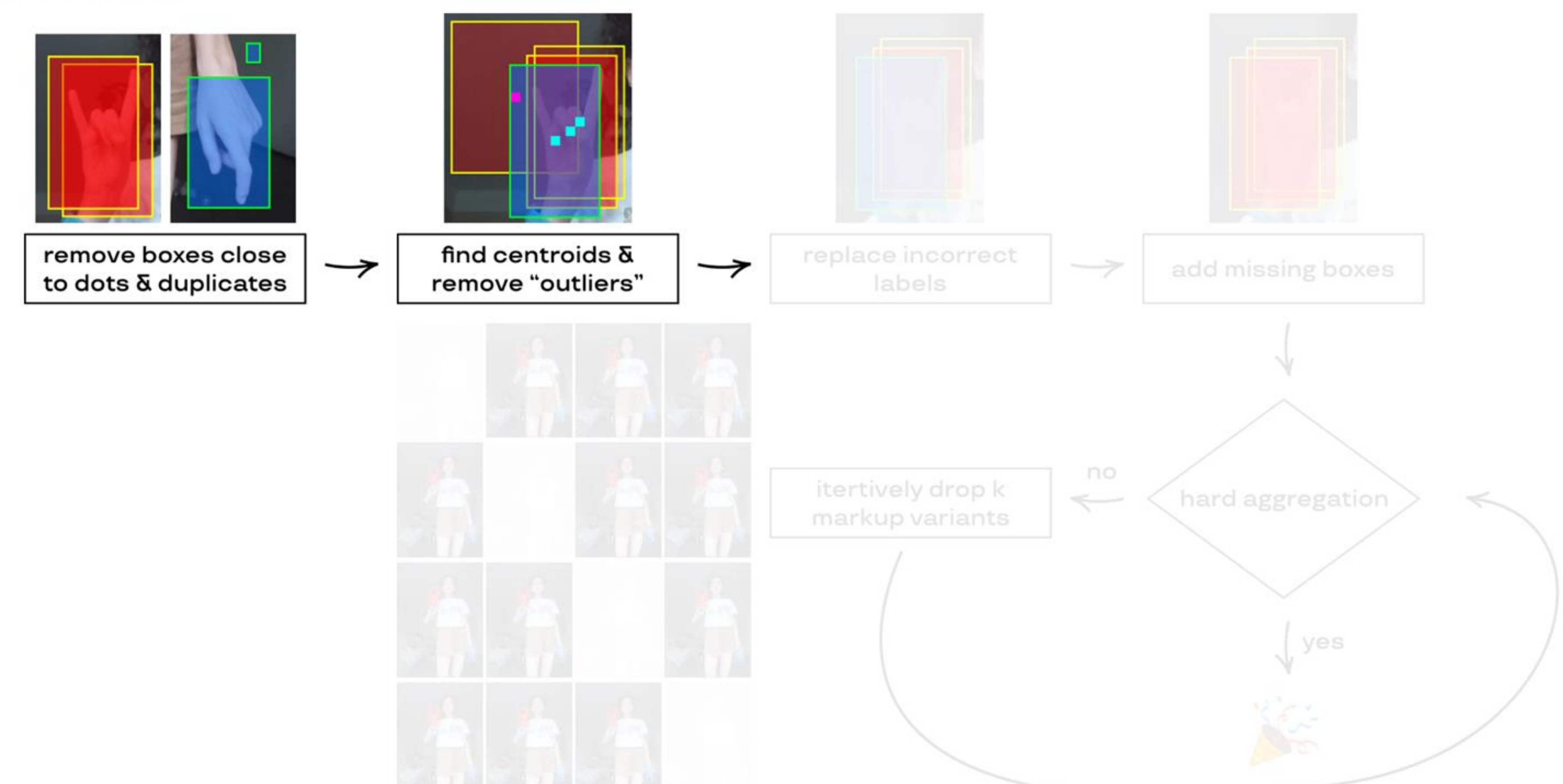






Агрегация разметки

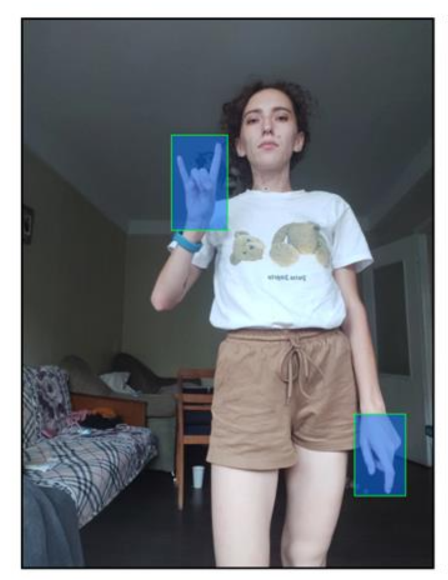
soft-подход

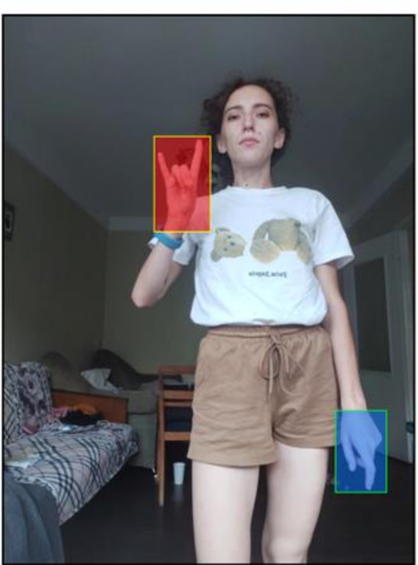


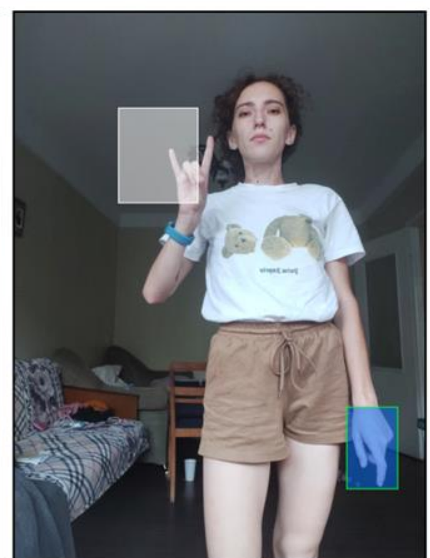
SBER DEVICES

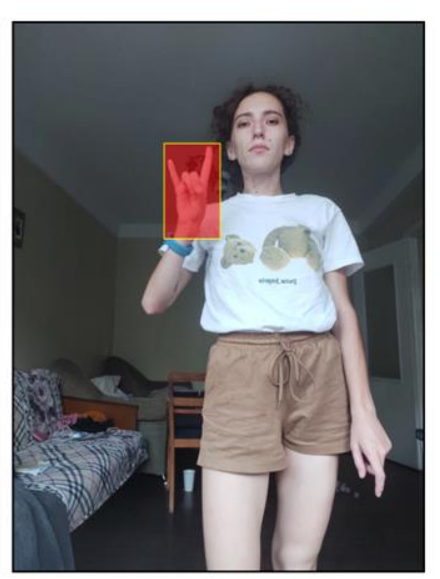
Агрегация разметки

soft-подход



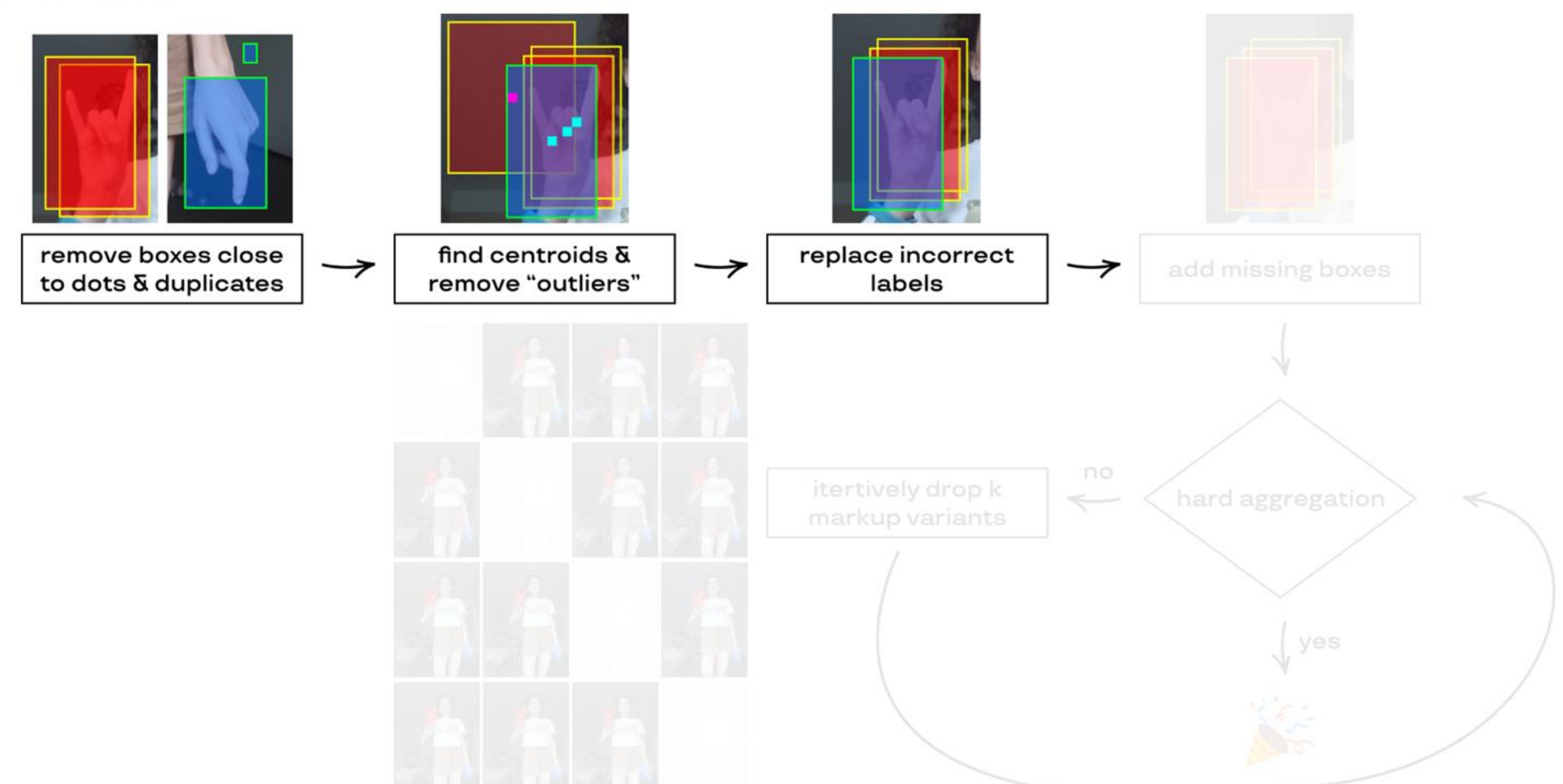






Агрегация разметки

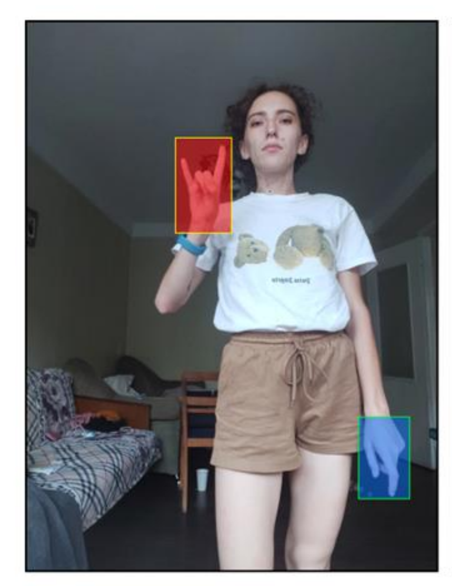
soft-подход

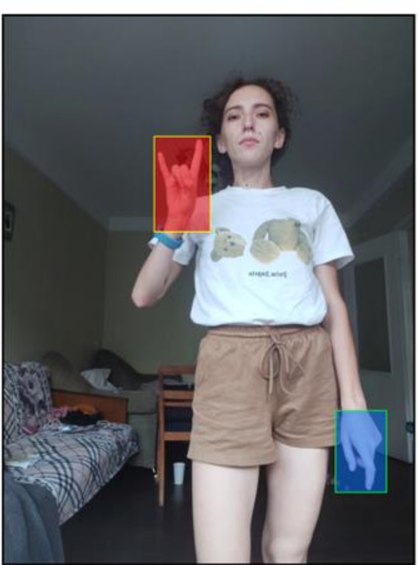


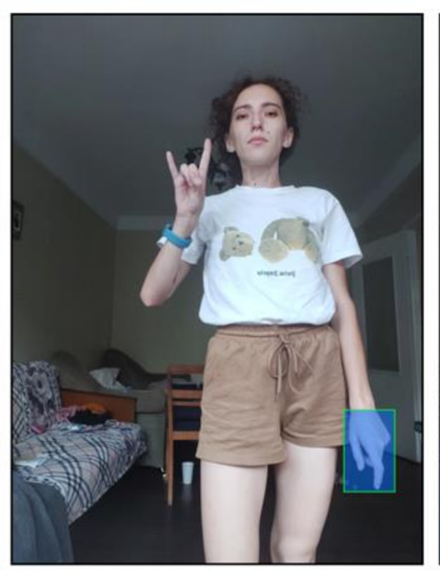
Агрегация разметки

soft-подход

SBER DEVICES



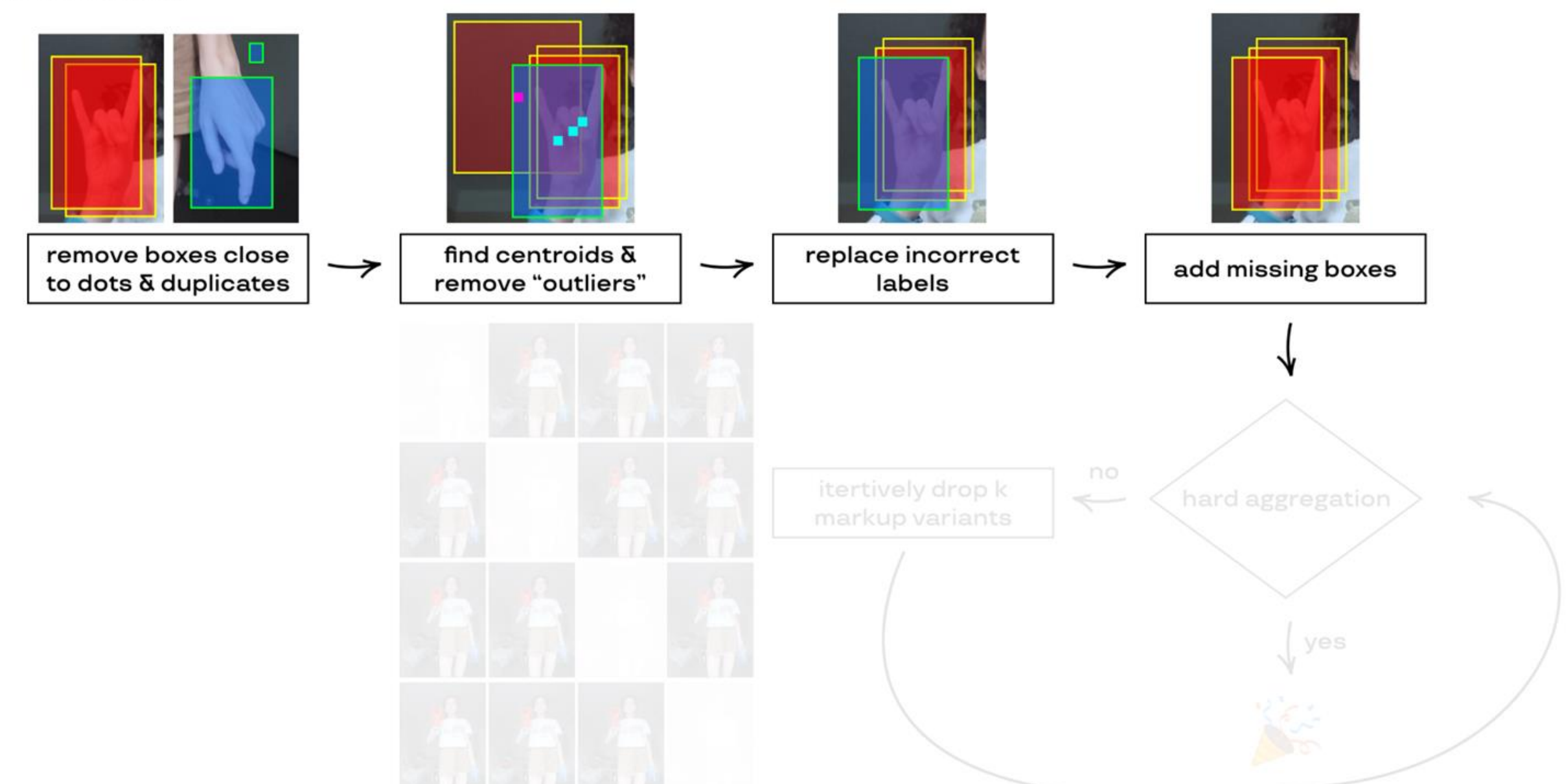




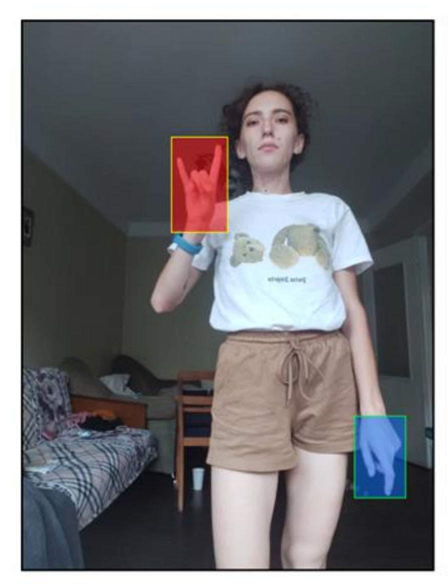
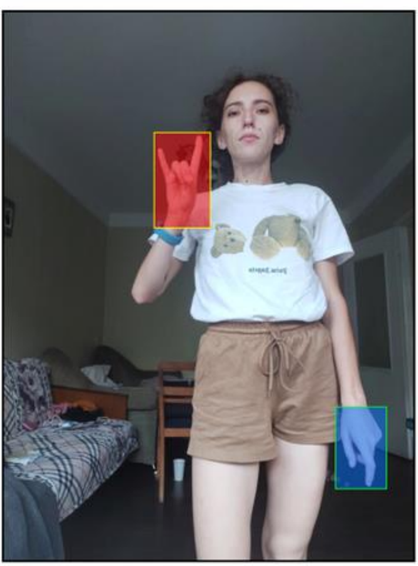
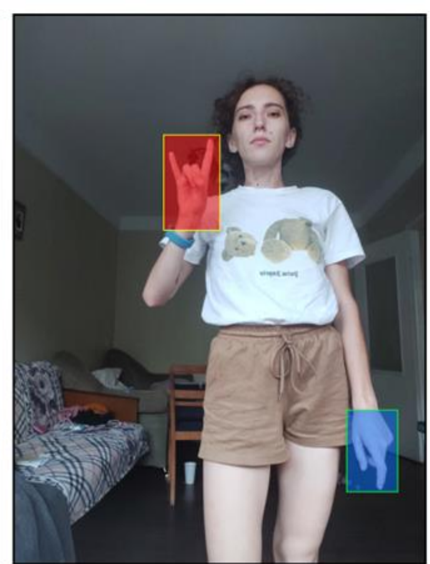
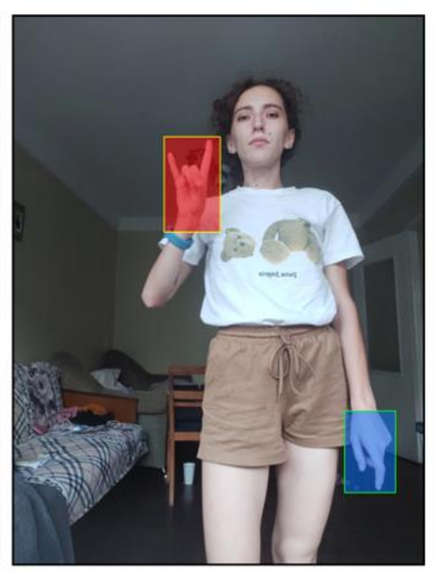


Агрегация разметки

soft-подход

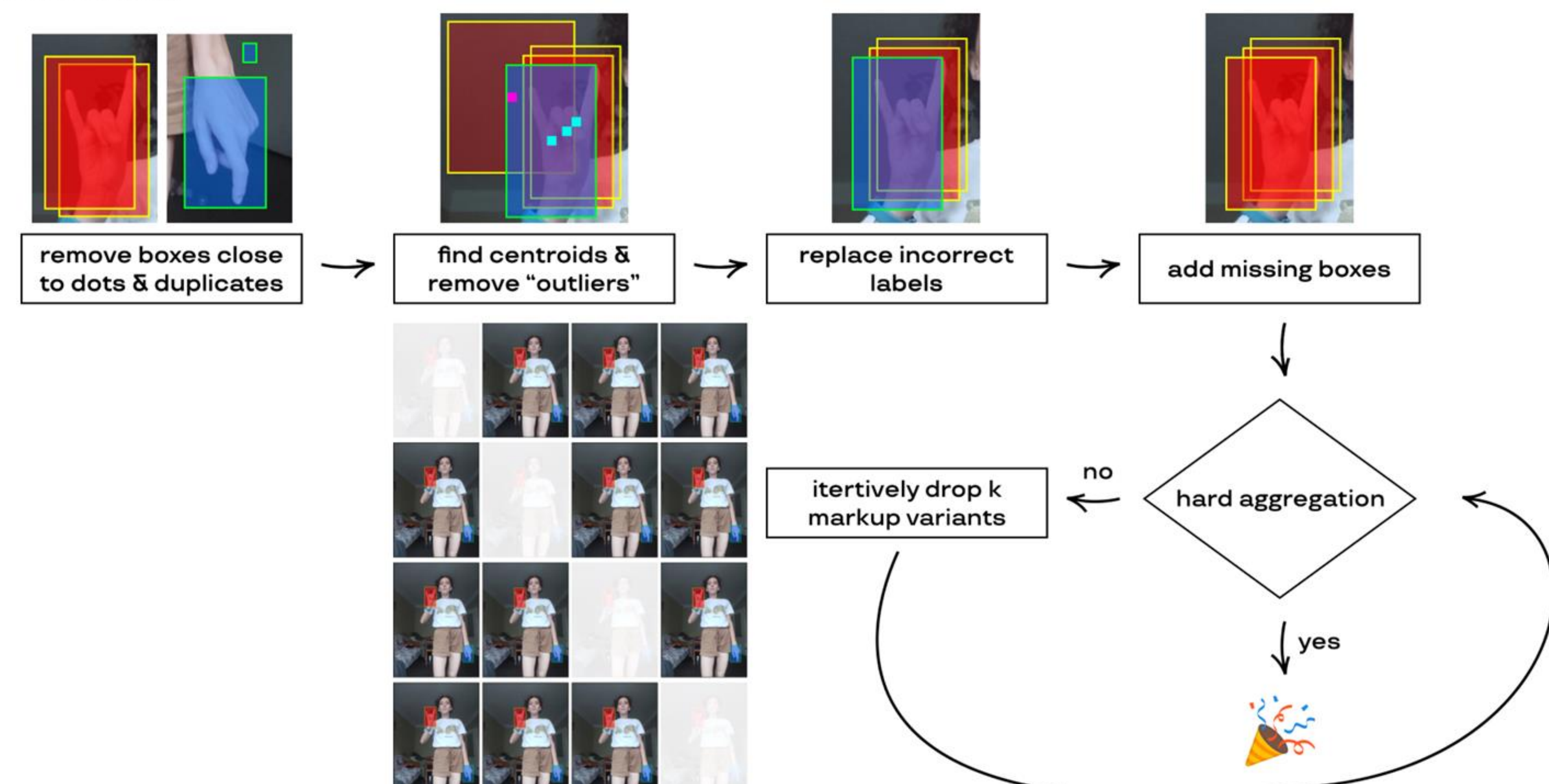


Агрегация разметки
soft-подход

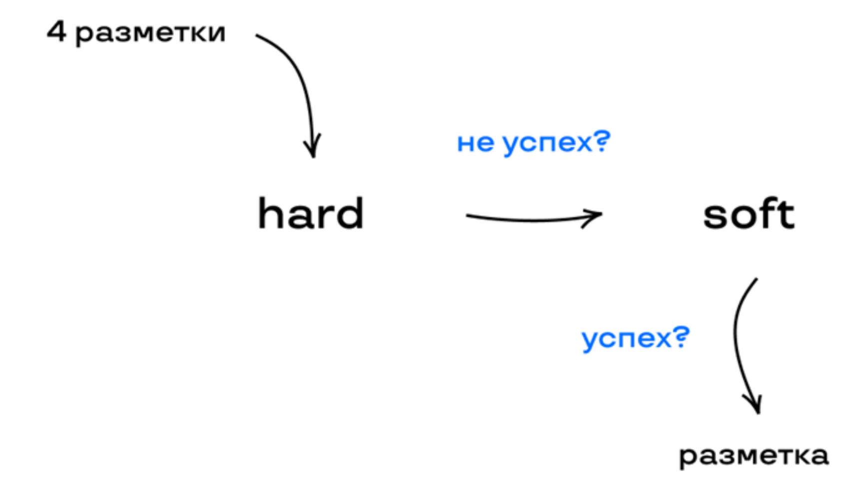





Агрегация разметки

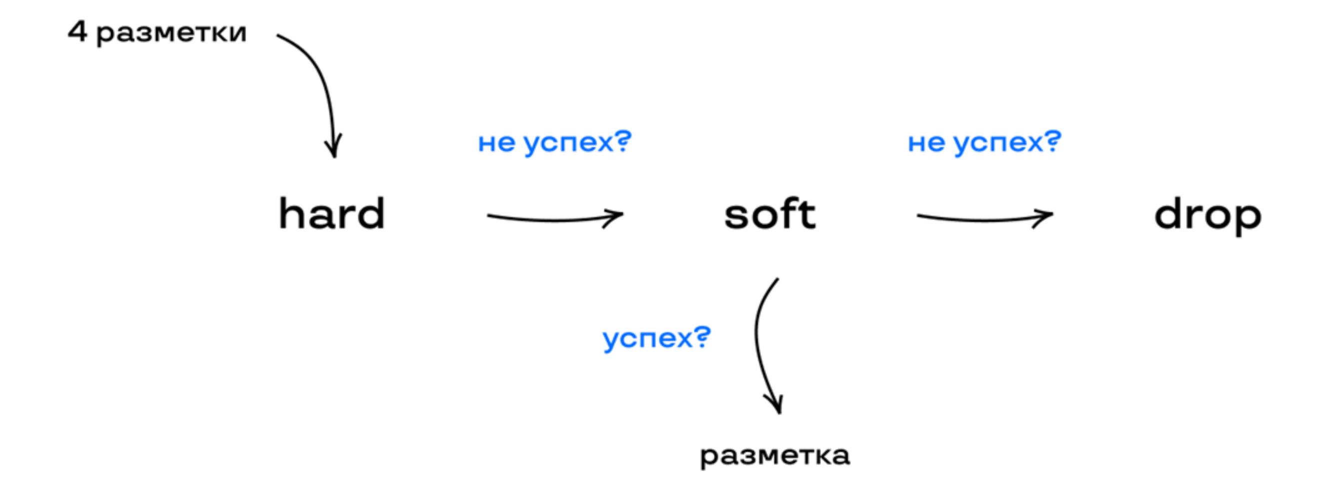
soft-подход



Агрегация разметки

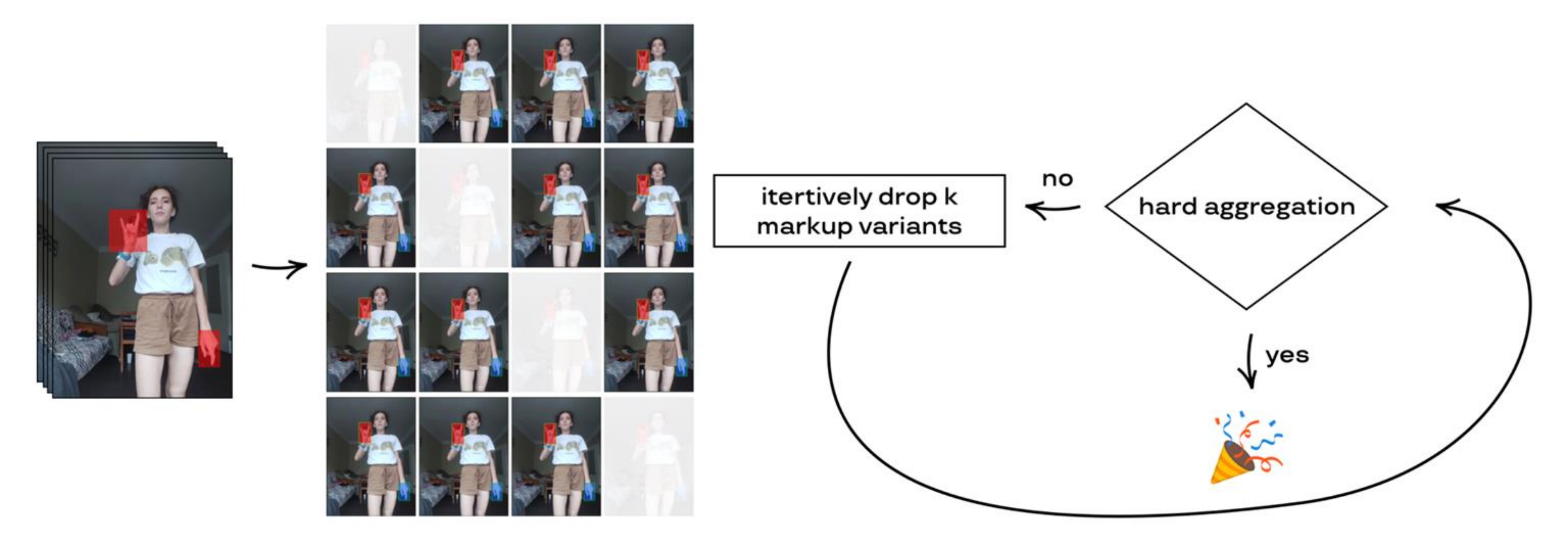


Агрегация разметки

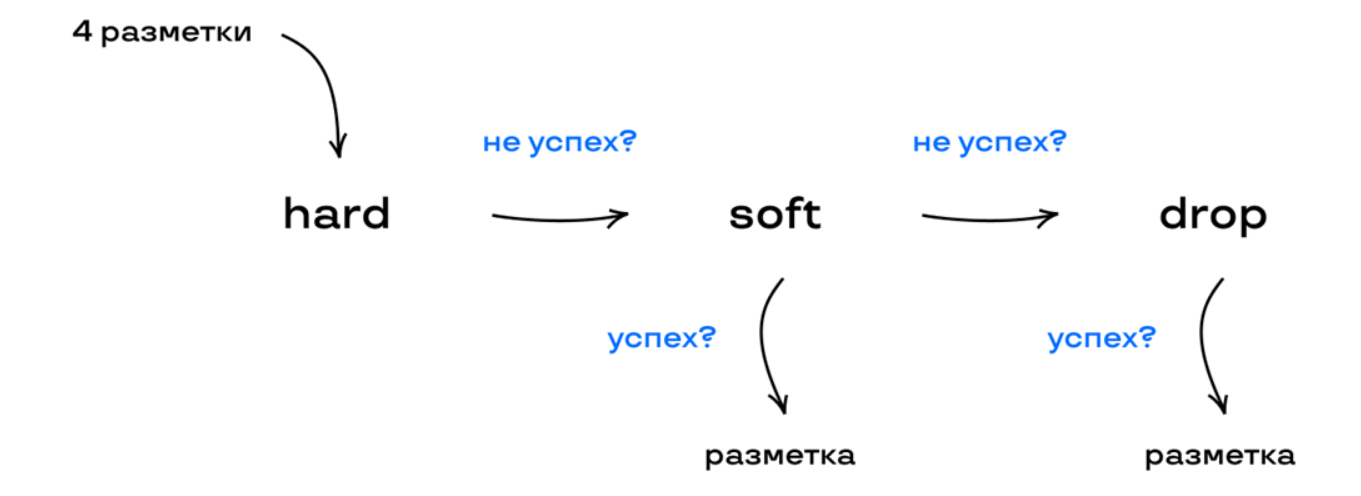


Агрегация разметки

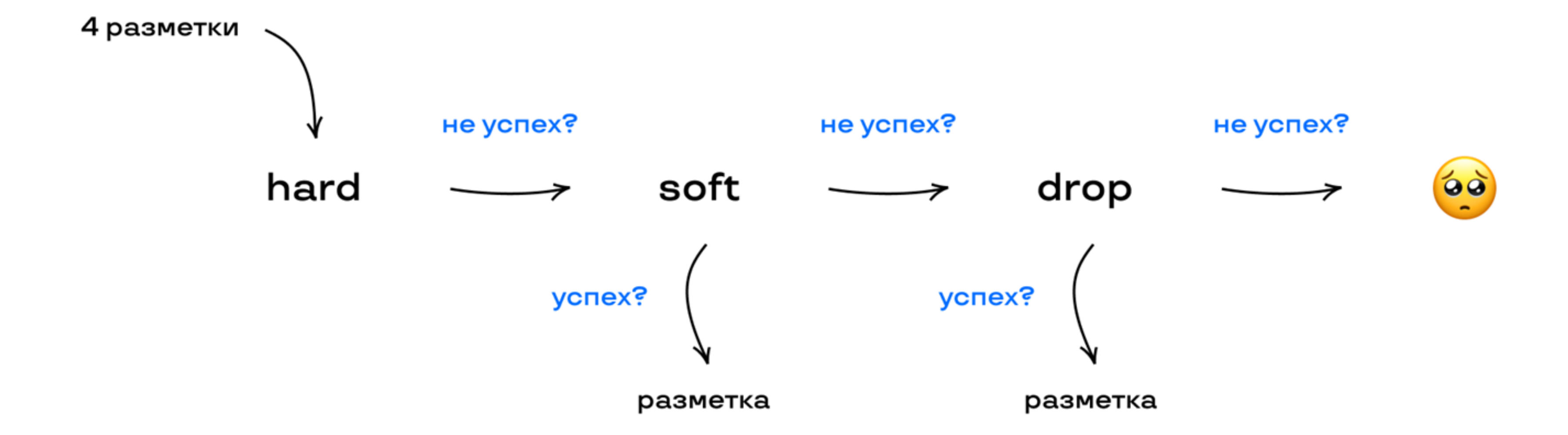
drop-подход



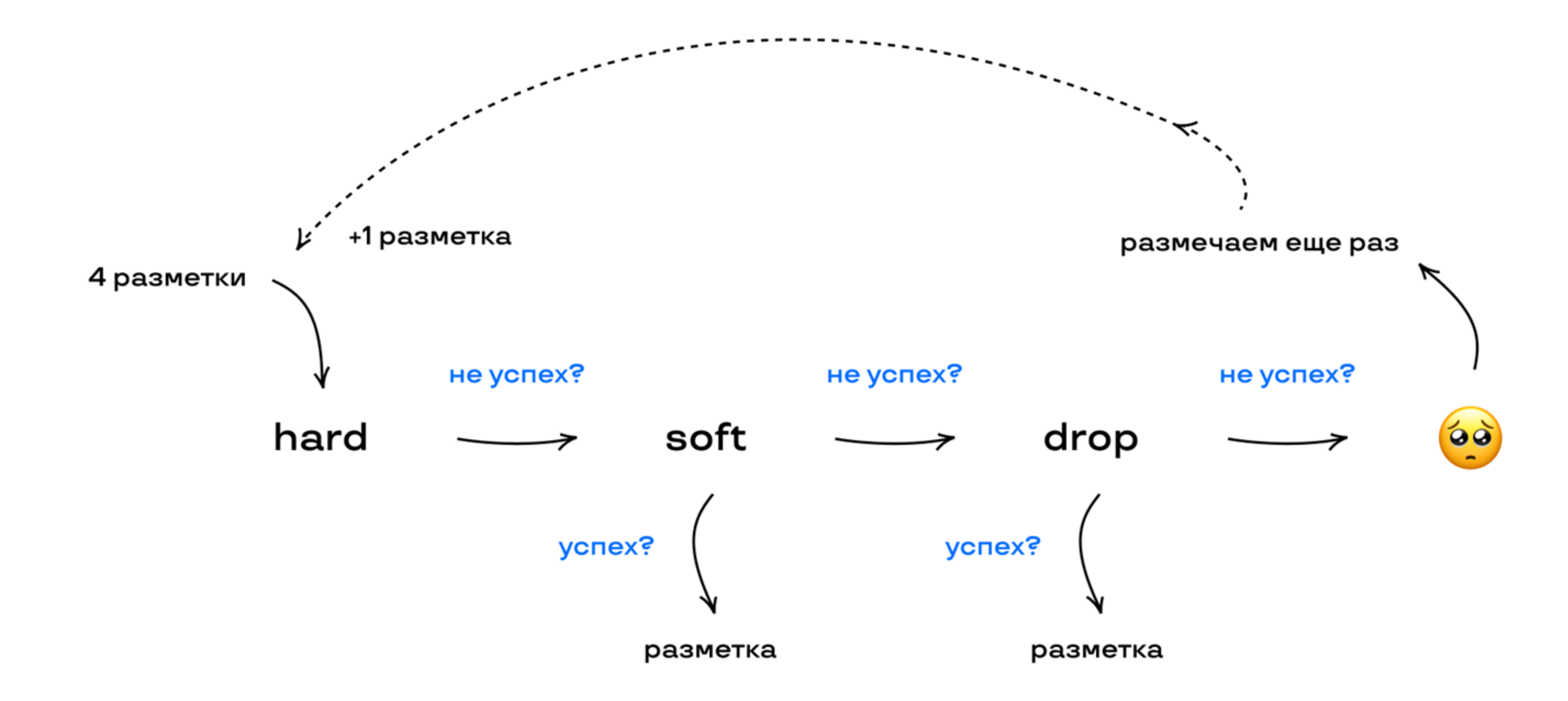
Агрегация разметки



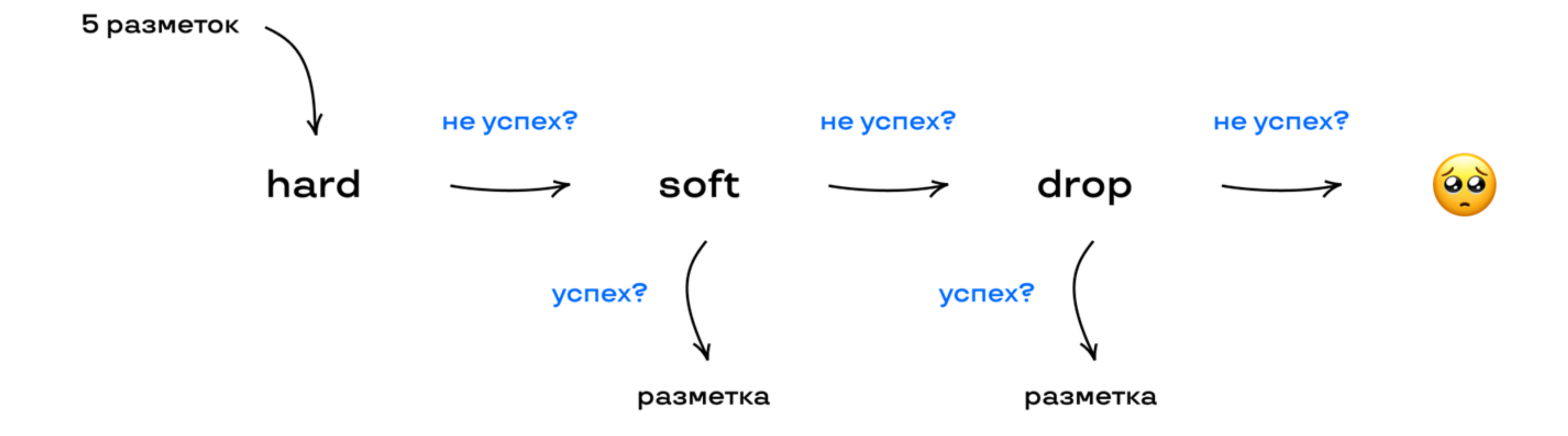
Агрегация разметки



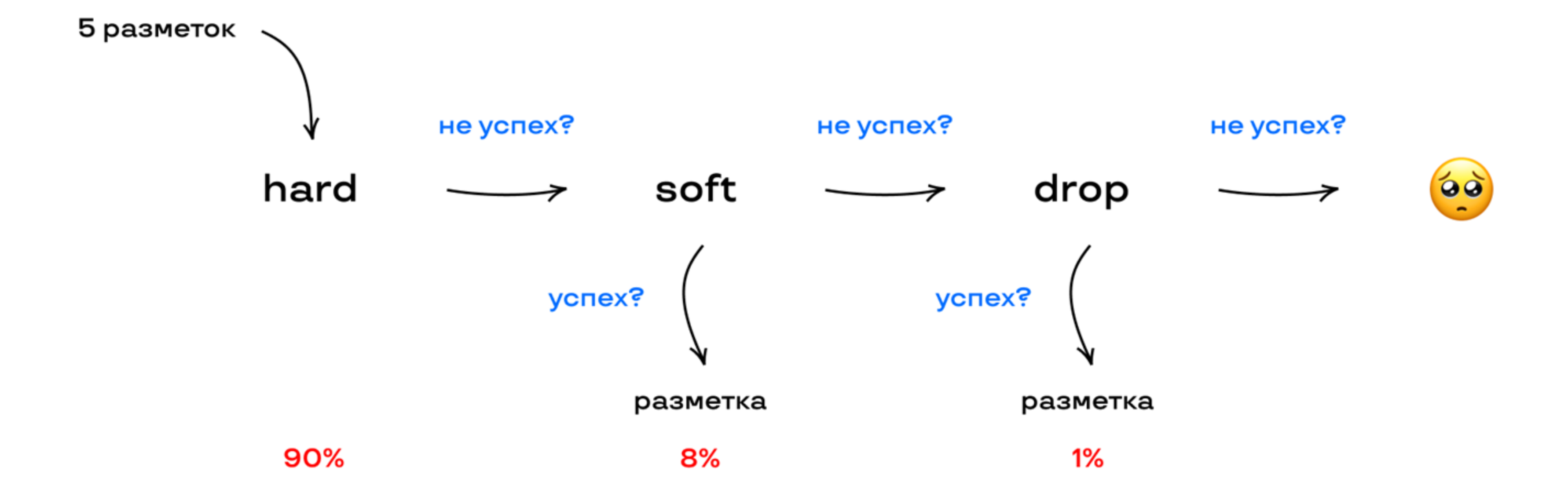
Агрегация разметки



Агрегация разметки



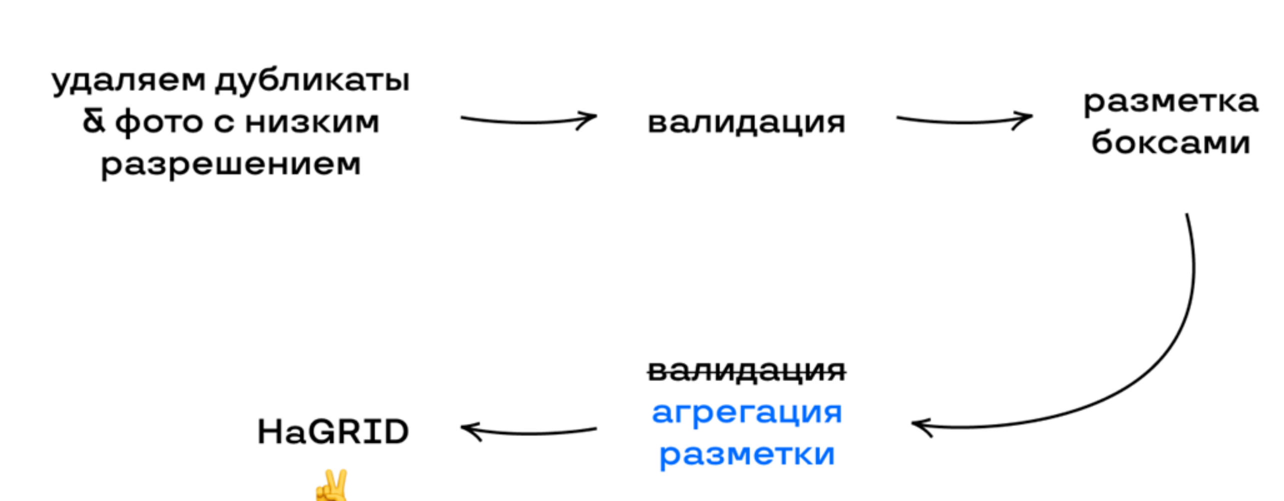
Агрегация разметки



Пайплайн сбора и разметки данных

сбор изображений





Стоимость пайплайна

каждое задание оценивалось в один цент (s = 0.01$)

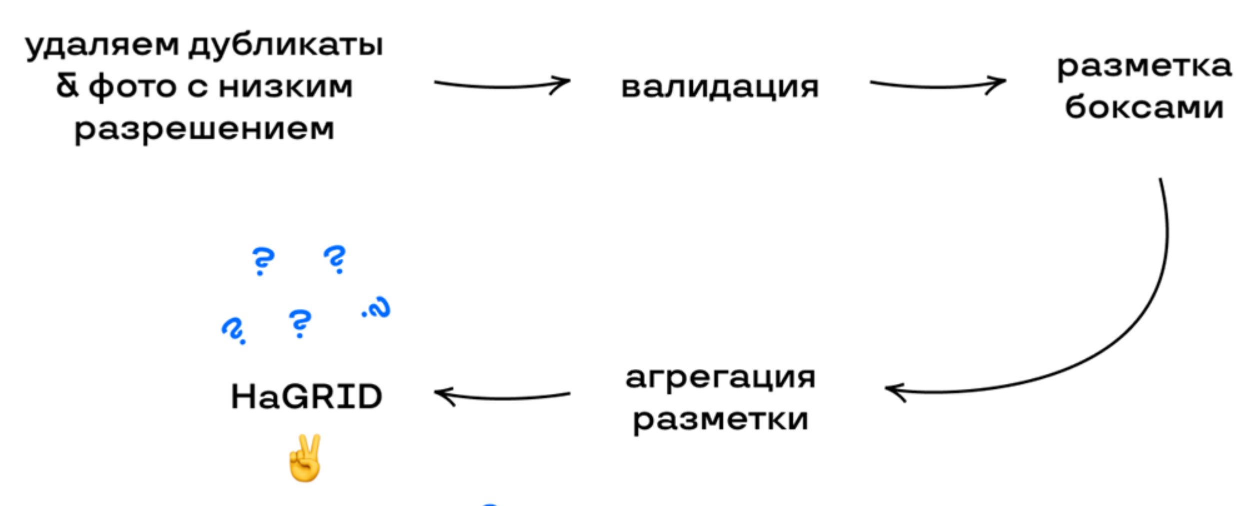
	n	по этапам	1 изображение
сбор изображений	878.910	n × s ≈ 9.000$	0.01$
валидация	830.679	n × s × 3 ≈ 25.000$	s × 3 ≈ 0.03$
разметка	560.230	n × s × 5 ≈ 28.000$	s × 5 ≈ 0.05$
агрегация разметки	560.230	140.000$ –> 0$	0.25$ –> 0$
классификация	552.992	n × s × 3 ≈ 17.000$	s × 3 ≈ 0.03$
итого	552.992	220.000$ –> 62.000$	0.37$ –> 0.09$

SBER DEVICES

Публикация в open-source

сбор изображений





Публикация в open-source

что стоит сделать для публикации в open-source?

Отсеять изображения

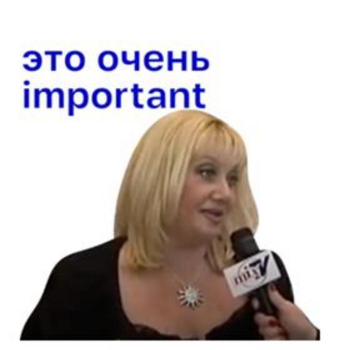




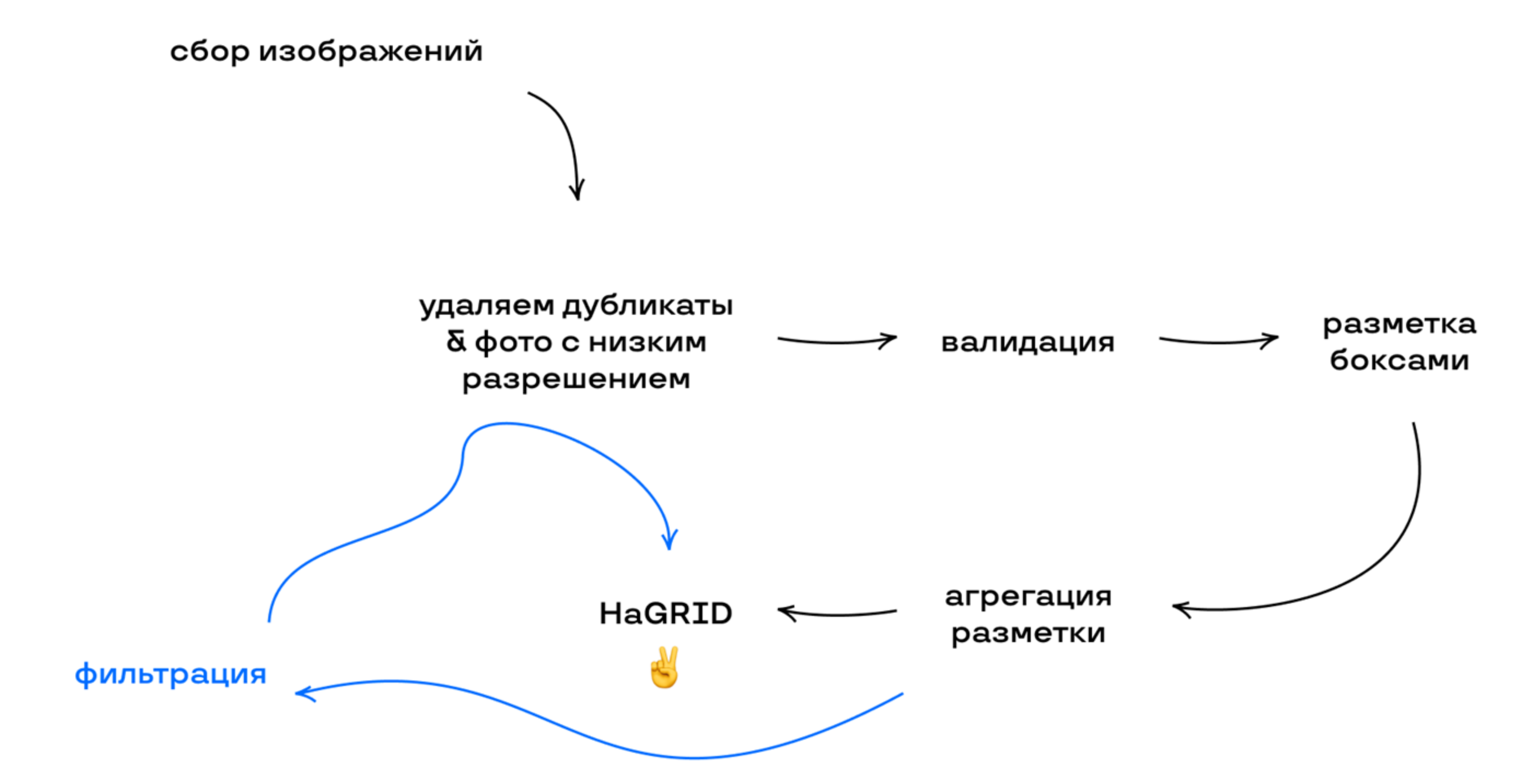


- с запрещенной символикой

это очень
important

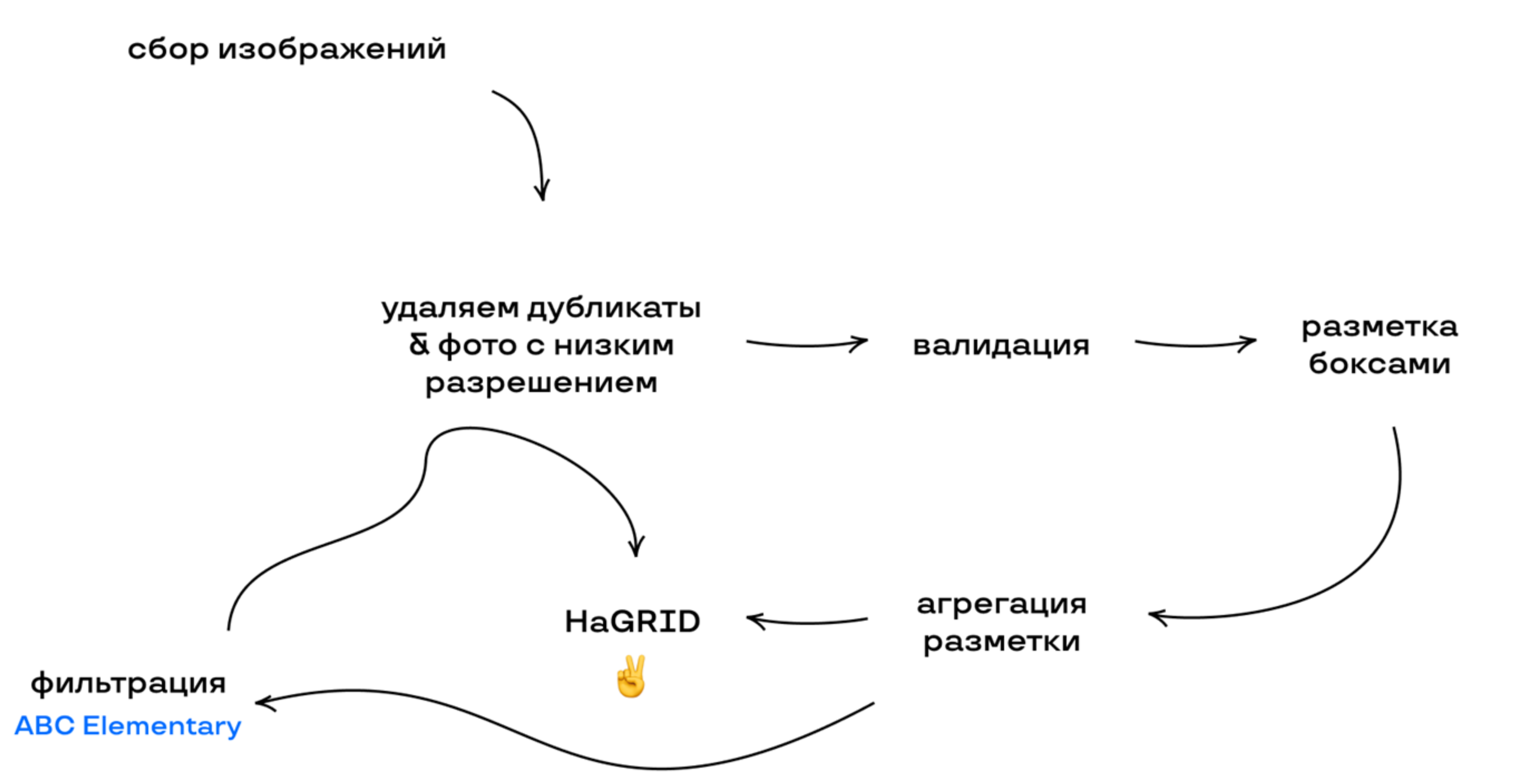


Публикация в open-source



SBER DEVICES

Публикация в open-source



Автоматизированный пайплайн

> python mining –p gestures –d image –c toloka –t <list_of_gestures> –overlap 1000



Название	Приоритет	Прогресс	Статус	Запущен
CALL [RNDCV] [Alexander Kapitanov]	0	8005* / 10000	Идёт разметка...	18 октября 2022 г., 9:22
LIKE [RNDCV] [Alexander Kapitanov]	0	6143* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
DISLIKE [RNDCV] [Alexander Kapitanov]	0	6134* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
FIST [RNDCV] [Alexander Kapitanov]	0	6785* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
STOP [RNDCV] [Alexander Kapitanov]	0	6645* / 9000	Идёт разметка...	18 октября 2022 г., 22:56
STOP_INVERTED [RNDCV] [Alexander Kapitanov]	0	1765* / 2000	Идёт разметка...	13 октября 2022 г., 13:20
THREE [RNDCV] [Alexander Kapitanov]	0	6273* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
TWO_UP [RNDCV] [Alexander Kapitanov]	0	712* / 1000	Идёт разметка...	18 октября 2022 г., 21:58
THREE2 [RNDCV] [Alexander Kapitanov]	0	681* / 1000	Идёт разметка...	18 октября 2022 г., 22:22







Автоматизированный пайплайн

какие еще данные собирали

skin segmentation

teeth segmentation

text to image

image to text

side by side

portrait segmentation

face detection

Ссылки



оценить доклад



Александр Капитанов
Team Lead CV RnD
@hukenovs

Карина Кванчиани
CV Engineer
@karinakv

arxiv



github



https://arxiv.org/abs/2206.08219

https://github.com/hukenovs/hagrid